Imputation-Powered Inference for Missing Covariates *

Junting Duan† Markus Pelger‡

This draft: March, 2026
First draft: November, 2025

Abstract

Missing covariate data is a prevalent problem in empirical research. We provide a novel framework for handling missing covariate data for estimation and inference in downstream tasks. Our general framework provides an automatic and easy-to-use pipeline for empirical researchers: First, missing values are imputed using a broad class of imputation methods under general observation patterns. Second, we automatically correct for the imputation bias and adaptively weight the imputed values according to their quality. Third, we use all available data, including imputed observations, to obtain more precise point estimates for the downstream task with valid confidence intervals. Our approach ensures valid inference while improving statistical efficiency by leveraging all available data. Through simulations, we demonstrate the superior performance of our approach over natural benchmarks, as it achieves both lower bias and variance while being robust to imputation quality. In a comprehensive empirical study of the dependence of equity markets on carbon emissions, we show that properly accounting for missing emissions data yields no evidence of correlation between stock returns and emissions directly produced by companies, but a negative correlation with value chain emissions.

Keywords: Missing Data, Imputation, Statistical Inference, Bias Correction, Adaptive Weighting, Method of Moments, Machine Learning, Panel Data, Stock Returns and Emissions

*We thank Will Cong, Kay Giesecke, Guido Imbens, Lihua Lei, Jann Spiess, Vasilis Syrkanis, Tijana Zrnic, and seminar and conference participants at Stanford, the University of Hong Kong, the National University of Singapore, the Chinese University of Hong Kong, Peking University, Tsinghua University, Fudan University, INFORMS, the ABFR Doctoral Research Symposium, and the World Online Seminar on ML in Finance. We thank Qube Research & Technologies for generous research funding and Brianna Xie for excellent research assistance.

†Stanford University, Department of Management Science & Engineering, Email: duanjt@stanford.edu.

‡Stanford University, Department of Management Science & Engineering, Email: mpelger@stanford.edu.

1 Introduction

Modern empirical research increasingly faces the problem of missing covariate data in large datasets. For example, in finance, key firm features such as environmental disclosures are frequently missing. In macroeconomics, different variables are collected at different frequencies, leading to structural missingness in low-frequency time series. In social sciences, survey-based covariates suffer from nonresponse or attrition, and administrative datasets may be incomplete or only partially observable. Missing covariates pose fundamental challenges for empirical analysis: when explanatory variables contain missingness, researchers cannot directly use them to estimate regression coefficients, test hypotheses, or identify causal relationships.

Despite its prevalence, the problem of missing covariates is often neglected in empirical research. In practice, researchers typically resort to two naive and problematic approaches as part of the data-cleaning process. The first is complete-case analysis, which discards all observations with any missing covariates and conducts analysis on the remaining fully observed subset. While simple to implement, this approach suffers from substantial data loss, reduced statistical efficiency, and potential selection bias. More severely, it becomes infeasible when there is no fully observed subset, a situation that is increasingly common in modern datasets with a large number of covariates. Alternatively, researchers may impute missing covariates and proceed as if the imputed values were actually observed. This second approach ignores imputation error, which can lead to biased estimates and invalid inference in subsequent estimation, producing spurious findings and misleading conclusions. These problems call for the need for a new principled solution for the missing covariate problem.

In this paper, we propose a novel framework for handling missing covariate data in downstream analysis. Our framework proceeds in three steps. First, we impute the missing values with a broad class of imputation methods, from matrix completion algorithms to flexible machine learning methods, under general observation patterns relevant in practice. The missingness can exhibit complex structures, including block and staggered patterns, and depend on the data through fully observed covariates. Second, given the imputed values, we automatically correct for the bias introduced by imputation and adaptively weight the imputed values based on their quality. Finally, we use all available data, including the imputed ones, in the downstream task, producing more precise point estimates with valid confidence intervals. This framework provides an automatic and easy-to-use pipeline for empirical researchers to estimate models from incomplete datasets. The process is simple: researchers input their incomplete dataset with complex observation patterns and choose their

preferred imputation method, and our framework provides the inferential theory for their target estimation.

Incorporating imputed data into downstream estimation raises several fundamental challenges. How do we account for the difference between truly observed data and imputed values? How do we remove the bias caused by general nonlinear imputation methods? How do we properly handle the data with heterogeneous observation patterns, where some units have less missingness and others have more? And how do we leverage high-quality imputations while protecting ourselves from poor ones? Our approach addresses these challenges based on two key ideas: correcting imputation bias using observed data and adaptively weighting imputed data based on their reliability for the downstream task.

To ensure unbiasedness when using imputation, we develop a data-driven debiasing method that quantifies and corrects the bias introduced by imputation using observed covariates. When fully observed units are available, they serve as natural benchmarks for bias correction. When no fully observed units exist, we instead construct debiasing terms by exploiting the relative completeness across partially observed units. For this data-driven debiasing to be valid, the imputation functions need to be independent of the data on which they are evaluated. We ensure this independence through a cross-fitting procedure, in which imputation models are trained on one subsample and applied to held-out data. Compared to closed-form bias corrections, our debiasing approach does not rely on parametric assumptions on the data-generating process and accommodates general nonlinear imputation methods.

To ensure efficient estimation, we adaptively weight the imputed values relative to observed ones. Specifically, we assign different weights to units with different observation patterns to account for heterogeneity in both the extent of missingness and imputation reliability. Observations with more extensive missingness and noisier imputations are downweighted. When the number of distinct observation patterns is large, we further introduce a data-driven grouping approach that clusters patterns with similar quality of imputed data and assigns each group a common weight. This grouping approach preserves the benefits of adaptive weighting while improving computational stability. When all partially observed units are assigned zero weight, our approach degenerates to the complete-case analysis using only the fully observed units. The optimal weights depend on imputation quality, the degree of missingness, and the signal-to-noise ratio of the estimation problem. They balance a fundamental efficiency trade-off: including more imputed data increases the effective sample size but also introduces additional noise. The resulting estimator is unbiased and efficient, leveraging all

available data.

We develop the asymptotic inferential theory for our estimator under general observation patterns (including heterogeneous and endogenous patterns) and with a broad class of imputation methods. This inferential theory serves two crucial purposes. First, it enables valid confidence intervals and hypothesis testing for downstream applications, ensuring safe deployment of imputation methods in high-stakes tasks. Second, it provides a data-driven solution to determine the optimal weights in our weighting scheme by minimizing the asymptotic variance of the estimator. We show that these weights solve a linear system of equations and automatically adapt to imputation quality.

Our framework is among the most general approaches for this problem in the literature, accommodating both general patterns of missing covariates and general imputation methods. Our solution remains useful even when the imputation method is misspecified or subject to large errors. In such cases, the framework adaptively reduces the weight on imputed values, ensuring performance that is no worse than the complete-case analysis. Given its generality, we expect this framework to be widely adopted across disciplines and fundamentally change how empirical researchers handle incomplete data.

We conduct extensive simulations to demonstrate the superior performance of our approach compared to the two commonly used but naive approaches: complete-case analysis and naive use of imputation. Our framework consistently achieves low bias and variance in all settings, while naive methods are generally biased, inefficient, or both. Compared to complete-case analysis, our approach achieves greater efficiency by appropriately incorporating information from partially observed units. Compared to the naive use of imputation, it demonstrates remarkable robustness to misspecified imputation models and poor imputation quality. When imputation accuracy is low, the adaptive weighting scheme downweights unreliable imputations and maintains strong performance in both bias and variance. Moreover, our simulation results show that heterogeneous weights play a crucial role when observation patterns exhibit large heterogeneity, and that a moderate number of groups is sufficient to capture most of the efficiency gains from group-specific weighting.

In our empirical study, we use our framework to answer a fundamental question in environmental finance: whether firms' carbon emissions correlate with their stock returns after controlling for firm characteristics (Bolton and Kacperczyk, 2021; Aswani, Raghunandan, and Rajgopal, 2024). The prior literature has yielded contradictory findings because of the extensive missingness of emission data. Previous studies find no correlation when using only the subset of firms with complete observations, but positive correlations when using vendor-imputed emission data. Using our framework, we find

that emissions from company operations are not correlated with stock returns; however, value chain emissions are negatively correlated with returns. These results demonstrate how methodological choices in handling missing data can fundamentally change empirical conclusions and highlight the importance of our framework to obtain reliable insights in the presence of missing covariates.

The remainder of the paper is organized as follows. Section 2 develops intuition for our approach through a simple linear regression example. Section 3 introduces the model setup and presents our general estimation framework. Section 4 establishes the asymptotic properties of our estimator and derives optimal weights that minimize the asymptotic variance. Section 5 studies endogenous missingness that arises through observed covariates and introduces a reweighted variant of our approach. Section 6 extends the framework to settings without fully observed units. Section 7 demonstrates the performance of our method through simulation studies, and Section 8 applies our framework to examine the relationship between carbon emissions and stock returns. Section 9 concludes. The Appendix contains supplementary methodological and implementation details, as well as the technical proofs and additional results.

1.1 Related Literature

Our work contributes to three streams of literature. First, we advance the literature on handling missing covariates in estimation problems by proposing a general framework that accommodates a broad class of imputation methods and general missing data patterns, while providing rigorous inferential theory for our estimator. Second, we extend recent advances in efficient estimation with missing outcomes to the more complex challenge of multivariate missing covariates. We introduce an adaptive weighting scheme that accounts for heterogeneous imputation reliability across different observation patterns. Third, we bridge the gap between the literature on missing data imputation methods and their application in downstream analysis, ensuring that imputed values can be safely incorporated into statistical inference. In summary, our framework provides the most general and precise approach to conduct inference with missing covariate data for downstream applications.

Our work fundamentally advances beyond the literature on handling missing covariates in downstream estimation (see, e.g., Little (1986)). Among existing methods that provide valid inferential theory with imputed values, Abrevaya and Donald (2017) develop an efficient estimator for linear regression models with linear imputation, formulated in a generalized method of moments (GMM) framework (Hansen, 1982). Freyberger, Höppner, Neuhierl, and Weber (2024) extend this by considering multiple observation patterns and propose a weighted least squares estimator, cast in a

GMM framework, that accounts for the additional variance introduced by imputation. While these approaches typically rely on correctly specified models and imputation methods, our framework does not impose parametric data-generating assumptions and accommodates general imputation methods within a unified inferential framework that allows for valid inference.[1]

Our framework pioneers a new research direction by extending recent advances in efficient estimation with missing outcomes to the more complex challenge of multivariate missing covariates. Our debiasing approach is motivated by the prediction-powered inference (PPI) framework developed by Angelopoulos, Bates, Fannjiang, Jordan, and Zrnic (2023), which uses fully observed data to correct the bias in predictions for missing outcomes.[2] However, our work initiates a new research direction in which missingness occurs in multivariate covariates (rather than the one-dimensional outcome) and accommodates multiple, heterogeneous observation patterns. Moreover, the framework can be applied even when no fully observed units are available.

Finally, our work bridges the gap between the literature on missing data imputation and its application to downstream analysis. A large body of literature studies imputation of missing entries in high-dimensional panels, including matrix completion and factor-based approaches.[3] These methods are designed to improve imputation accuracy as a standalone objective and do not address how the imputed values should be properly incorporated into downstream analyses. In contrast, our work focuses on this subsequent step and provides a framework for using imputed values to obtain valid and efficient statistical conclusions.

The inferential theory for imputation-based estimators with missing data is an active area of

[1]The widely used missing indicator method (MIM) in the machine learning literature is generally inconsistent, can introduce bias in downstream estimation, and does not guarantee efficiency improvements over complete case analysis (Jones, 1996; Donders, van der Heijden, Stijnen, and Moons, 2006; Van Ness, Bosschieter, Halpin-Gregorio, and Udell, 2023).

[2]A growing body of research in semi-supervised learning studies how to leverage large amounts of data with missing outcomes to improve estimation when the number of fully observed data is small (Wasserman and Lafferty, 2007; Chakrabortty and Cai, 2018; Deng, Ning, Zhao, and Zhang, 2024). Among these works, Wang, McCormick, and Leek (2020) and Angelopoulos, Bates, Fannjiang, Jordan, and Zrnic (2023) consider settings where a prediction model for outcomes is available and propose methods that correct prediction errors for valid downstream inference. Recent extensions of the PPI framework include Angelopoulos, Duchi, and Zrnic (2023) and Miao, Miao, Wu, Zhao, and Lu (2023), which develop methods that adapt to the precision of machine learning predictions; Ji, Lei, and Zrnic (2025), which proposes a recalibration approach to learn the optimal loss for data with predicted labels; and Kluger, Lu, Zrnic, Wang, and Bates (2025), which develops bootstrap-based approaches to construct confidence intervals for predict-then-debiased estimators.

[3]The problem of imputing missing entries in large panels has been extensively studied in the matrix completion literature since Candès and Recht (2012). Matrix completion methods typically exploit the low-rank structures of the underlying data matrix using nuclear norm regularization (Mazumder, Hastie, and Tibshirani, 2010; Negahban and Wainwright, 2011; Chen, Fan, Ma, and Yan, 2019; Chernozhukov, Hansen, Liao, and Zhu, 2023; Choi and Yuan, 2024). Recent approaches leverage latent factor models to provide inferential theory for imputed values (Jin, Miao, and Su, 2021; Bai and Ng, 2021; Xiong and Pelger, 2023; Cahan, Bai, and Ng, 2023; Duan, Pelger, and Xiong, 2024a,b; Bryzgalova, Lerner, Lettau, and Pelger, 2025).

research. Our paper is most closely related to recent contemporaneous and independent work by Zhao and Candès (2025) and Chen, McCormick, Mukherjee, and Wu (2025), which, together with our paper, develop inference procedures that combine imputation with explicit debiasing and reweighting of imputed values. These papers address a common goal from complementary methodological perspectives and differ in the generality of missing patterns that can be accommodated.[4]

Independent work by Zhao and Candès (2025) constructs their estimator via a one-step Newton update based on a weighted debiased loss, starting from a consistent initial estimator, whereas we directly formulate and solve the corresponding weighted debiased moment equations. They relax their missing-completely-at-random assumption to subpopulation inference under a testable first-moment missing-completely-at-random condition, which requires that, for each partially observed pattern, the imputed score function has the same expectation as that based on the fully observed data. Chen, McCormick, Mukherjee, and Wu (2025) propose an estimator that combines pattern-specific parameter estimators computed on data with imputed values together with their corresponding debiasing terms. They propose matrix-valued weights for the parameter estimator associated with each partially observed pattern. In comparison, we construct a debiased moment estimator and use weights for the moment estimator under each partially observed pattern. They adapt their method to settings with known or correctly specified propensity scores, allowing the observation pattern to depend endogenously on the data through observed covariates. Both papers provide important insights into how debiasing and reweighting can be combined with imputation to enable valid inference.

Our contribution is complementary to these approaches. We allow more general observation patterns than the other approaches. More specifically, we consider settings with a large number of distinct observation patterns. We introduce a data-driven procedure that groups observation patterns with similar imputation quality and assigns each group its own optimal weight, allowing adaptive reweighting to remain stable when the number of distinct patterns is very large. Importantly, we propose an approach that can be applied even in the absence of fully observed units. This setting is increasingly common as modern datasets include richer covariate sets, making it more likely for each unit to be missing at least one covariate and leaving no fully observed units. These features substantially broaden the range of empirical settings in which our approach can be applied.

[4]Other recent papers, including Berrett (2024) and Xu, Testa, Lei, and Roeder (2025), depart further from our framework by adopting a semiparametric efficiency perspective and constructing estimators via approximations to the efficient influence function.

2 Illustrative Example

We begin with a simple linear regression example to illustrate the fundamental challenges of missing covariates in model estimation and demonstrate the key insights of our proposed method before presenting our general framework.

2.1 Simple Setup

We observe a data set with N i.i.d. outcome-covariate pairs (Y_i, X_i), where $Y_i \in \mathbb{R}$ denotes a scalar outcome and $X_i \in \mathbb{R}^L$ denotes an L-dimensional covariate vector. In this simple illustrative example, we consider the problem of estimating the linear regression coefficient $\theta_0 \in \mathbb{R}^L$, defined as the unique solution to the moment condition

$$G(\theta) = \mathbb{E}[X_i X_i^\top \theta - X_i Y_i] = 0.$$

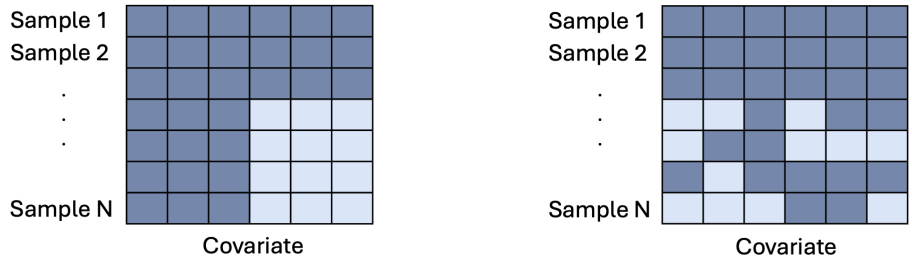
We focus on settings where the outcomes Y_i are fully observed, but the covariate vectors X_i contain missing entries. This situation is prevalent in practice, particularly when L is large. Throughout this example section, we assume that the observation pattern is exogenous and independent of the data. For illustration, we consider the simplest setting where N_0 units have all covariates fully observed, while the remaining $N_1 = N - N_0$ units share a single partially observed pattern: they observe only the first $L_0 < L$ covariates, with the remaining $L - L_0$ entries missing. Figure 1(a) illustrates this pattern. The key challenge in this setting is that the partially observed N_1 units contain valuable information, but cannot be directly incorporated into standard estimation procedures. To address this issue, empirical researchers typically resort to two approaches that either discard these partially observed units or impute the missing covariates and treat the imputed data as if they were truly observed, as part of a data pre-processing or cleaning step.

In the following, we first review these commonly used approaches, highlight their limitations, and then introduce our approach that addresses them. The insights from this example extend naturally to the general framework in Section 3.

2.2 Naive Approaches and Limitations

The most widely used approach for handling missing data is the *complete-case analysis*, which discards all observations with any missing entries and estimates the parameter using only the fully observed units. In this simple illustrative example, the complete-case estimator of θ_0 is obtained by

Figure 1: Examples of covariate observation patterns



(a) Homogeneous missing pattern (b) Heterogeneous missing patterns

These figures show examples of observation patterns for the covariate panel $\mathbf{X} = [X_1, \ldots, X_N]^\top \in \mathbb{R}^{N \times L}$. Each row represents a unit and each column represents a covariate. The dark color indicates observed entries, while the light color indicates missingness. Figure (a) shows a homogeneous missing pattern where all partially observed units are missing the same variables (the last $L - L_0$ covariates). Figure (b) shows heterogeneous missing patterns where each partially observed unit can have a different subset of missing variables.

solving the sample moment condition

$$G_N^{\text{complete}}(\theta) = \frac{1}{N_0} \sum_{i=1}^{N_0} \left(X_i X_i^\top \theta - X_i Y_i \right) = 0.$$

While simple to implement, this approach discards information contained in the partially observed units. As a result, it can suffer from substantial efficiency loss when missingness is prevalent ($N_0 \ll N$), leading to large standard errors. Moreover, this approach becomes infeasible in the case where every unit contains at least one missing covariate, a setting we address in Section 6.

An alternative approach, *naive imputation*, fills in missing values through imputation and proceeds with standard estimation procedures as if the imputed values were truly observed. For a partially observed unit with observation pattern $W = [\mathbf{1}_{L_0}; \mathbf{0}_{L-L_0}] \in \{0,1\}^L$ (where $W_l = 1$ if the l-th covariate is observed and $W_l = 0$ if missing), an imputation function takes both the covariate vector and its observation pattern as inputs and maps them to an imputed covariate vector $\tilde{X}$:

$$f : (X, W) \mapsto \tilde{X}, \quad \text{where } \tilde{X}_l = \begin{cases} X_l, & \text{if } W_l = 1 \\ \hat{X}_l, & \text{if } W_l = 0 \end{cases},$$

where $\hat{X}_l$ denotes an imputed value, constructed either using simple rules such as mean or median imputation or using more flexible data-driven methods that condition on the observed entries of X.

Giving the imputed covariate vectors $\tilde{X}_i$, the naive imputation estimator treats these values as if

they were observed and solves

$$G_N^{\text{Naive Imp}}(\theta) = \frac{1}{N}\left(\sum_{i=1}^{N_0}(X_i X_i^\top \theta - X_i Y_i) + \sum_{i=N_0+1}^{N}(\tilde{X}_i \tilde{X}_i^\top \theta - \tilde{X}_i Y_i)\right) = 0.$$

This approach uses all available data but ignores the uncertainty of the imputation.

Since imputation is rarely perfect, treating imputed values as truth can lead to biased estimates and invalid inference. For example, to impute a missing covariate $X_{i,l}$, standard imputation methods typically target the conditional mean $\mathbb{E}[X_{i,l} \mid \mathbf{X}_{i,\text{obs}}]$ based on the observed covariates. However, the linear regression moment above depends on second-order quantities, and targeting the conditional mean does not generally yield accurate approximations to the second moments required for estimation. As a result, even when the imputation is unbiased, substituting imputed covariates into the regression moment can lead to biased estimation.

2.3 Our Solution

Our framework addresses the limitations of these naive approaches by incorporating partially observed units with imputed covariates into the estimation while explicitly accounting for the errors introduced by imputation. This method is built on two key ideas: (1) correcting imputation-induced bias using the observed data, and (2) adaptively weighting the imputed values based on their reliability. These innovations work together to efficiently incorporate the information from all available data, while ensuring valid inference.

The first key idea is that the observed data provide a natural benchmark for quantifying the imputation error. For the fully observed units, we can compute the moment conditions using both the observed covariates and their imputed counterparts, thereby estimating the bias introduced by the imputation. In the illustrative example, this bias can be measured empirically on the N_0 fully observed units and used to construct a debiased moment estimator

$$G_N(\theta) = G_N^{\text{Naive Imp}}(\theta) + \underbrace{\frac{N_1}{N} \cdot \frac{1}{N_0}\sum_{i=1}^{N_0}\Big((X_i X_i^\top \theta - X_i Y_i) - (\tilde{X}_i \tilde{X}_i^\top \theta - \tilde{X}_i Y_i)\Big)}_{\text{debiasing term}},$$

which corrects for the bias induced by naively treating the imputed values as observed.

The second idea concerns the reliability of imputed values. Even after debiasing, the imputed covariates $\tilde{X}_i$ differ from the observed ones and should not necessarily contribute equally to the

estimation. To account for this heterogeneity, we introduce a weight parameter λ that controls the contribution of the imputed data.[5] Combining this weighting with bias correction yields the following weighted debiased moment estimator

$$G_N^\lambda(\theta) = \frac{1}{N_0}\sum_{i=1}^{N_0}(X_iX_i^\top\theta - X_iY_i) + \lambda\cdot\frac{N_1}{N}\left(\frac{1}{N_1}\sum_{i=N_0+1}^{N}(\tilde{X}_i\tilde{X}_i^\top\theta - \tilde{X}_iY_i) - \frac{1}{N_0}\sum_{i=1}^{N_0}(\tilde{X}_i\tilde{X}_i^\top\theta - \tilde{X}_iY_i)\right). \tag{1}$$

We define $\hat{\theta}^\lambda$ as the solution to $G_N^\lambda(\theta) = 0$.

The weight λ plays a central role in the estimation. When $\lambda = 0$, (1) degenerates to the complete-case analysis, using only the fully observed units. When the imputation is perfect (i.e., $\tilde{X}_i = X_i$), setting $\lambda = 1$ recovers the oracle moment function that would be available if all covariates were observed. More generally, the optimal choice of λ depends on the imputation quality: larger values are preferred when the imputation is accurate, while smaller values protect against poor imputation. In Section 4, we derive the optimal weight λ^* by minimizing the asymptotic variance of $\hat{\theta}^\lambda$, which balances the efficiency gain of including more data against the additional noise from them.

To further illustrate how λ affects the efficiency of the estimation, we analyze a simplified setting with a linear model $Y_i = X_i^\top\theta_0 + \epsilon_i$ with covariates $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_X)$ and noise $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. We abstract away from a specific imputation method by modeling the imputed values as noisy observations, i.e., the imputed values are $\hat{X}_l = X_l + \Delta_l$ with $\Delta_l \overset{i.i.d.}{\sim} \mathcal{N}(\mu_\Delta, \sigma_\Delta^2)$.

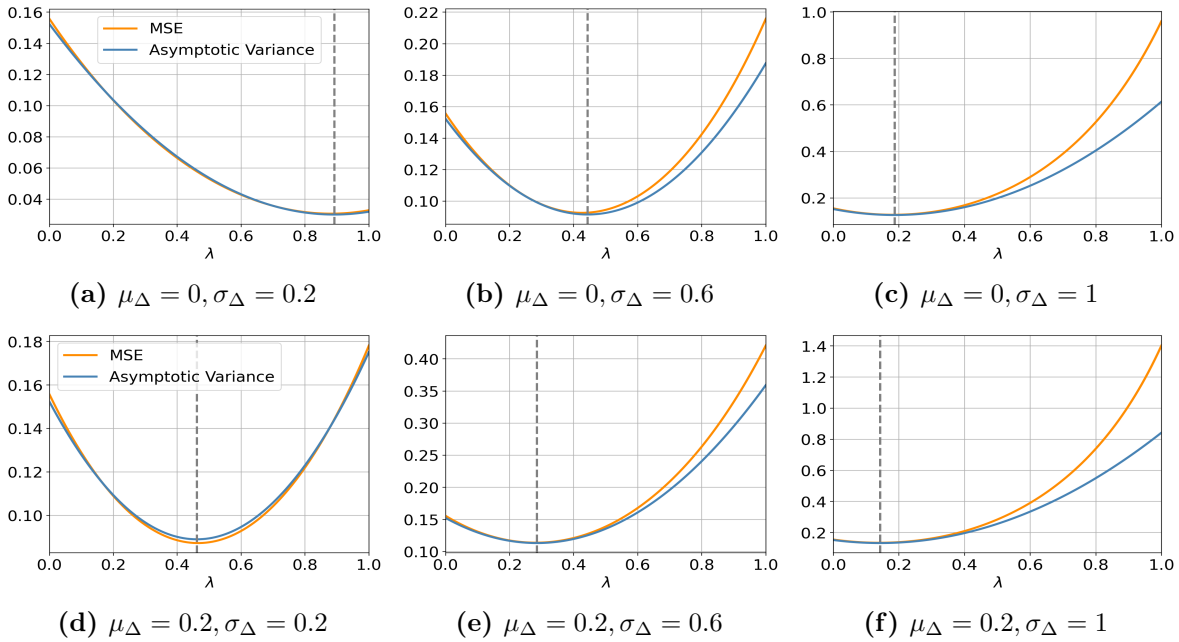
Proposition 1. *Let* $r_1 = \lim_{N\to\infty} N_1/N$, $\theta_0^{(1)} = \sum_{l=L_0+1}^{L}\theta_{0l}$ *and* $\theta_0^{(2)} = \sum_{l=L_0+1}^{L}\theta_{0l}^2$. *Suppose* $N_0 \to \infty$ *as* $N \to \infty$. *Under the data-generating process and observation pattern described above, assume* $\hat{\theta}^\lambda$ *uniquely solves* $G_N^\lambda(\theta) = 0$ *in* (1) *with probability tending to one, and that either (1) the parameter space* Θ *is compact, or (2)* $G_N^\lambda(\theta)$ *is nondecreasing, i.e.,* $(G_N^\lambda(\theta_1) - G_N^\lambda(\theta_2))^\top(\theta_1 - \theta_2) \geq 0$ *for any* $\theta_1, \theta_2 \in \Theta$. *Then, as* $N \to \infty$, *the estimator* $\hat{\theta}^\lambda$ *is asymptotically normal:*

$$\sqrt{N_0}(\hat{\theta}^\lambda - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma^\lambda),$$

with asymptotic variance $\Sigma^\lambda = \left[(1-2r_1\lambda)\sigma_\epsilon^2 + \lambda^2 r_1(\sigma_\epsilon^2 + \theta_0^{(2)}\sigma_\Delta^2 + (\theta_0^{(1)})^2\mu_\Delta^2)\right]\Sigma_X^{-1} + \lambda^2 r_1\Sigma_X^{-1}\Sigma_1\Sigma_X^{-1}$, *where* $(\Sigma_1)_{kl} = I(k,l > L_0)\big(\mu_\Delta^2\sigma_\epsilon^2 + \theta_{0k}\theta_{0l}\sigma_\Delta^4 + ((\theta_{0k}+\theta_{0l})\theta_0^{(1)} + \theta_0^{(2)})\mu_\Delta^2\sigma_\Delta^2\big) + I(k = l > L_0)\big(\sigma_\epsilon^2\sigma_\Delta^2 +$

[5] We recommend constraining the weight $\lambda \in [0, 1]$ for high-stakes applications. However, the optimal λ may exceed 1 when imputation provides beneficial regularization effects (e.g., filtering out covariate noise through imputation).

Figure 2: Estimation efficiency for different weights λ



(a) $\mu_\Delta = 0, \sigma_\Delta = 0.2$ **(b)** $\mu_\Delta = 0, \sigma_\Delta = 0.6$ **(c)** $\mu_\Delta = 0, \sigma_\Delta = 1$

(d) $\mu_\Delta = 0.2, \sigma_\Delta = 0.2$ **(e)** $\mu_\Delta = 0.2, \sigma_\Delta = 0.6$ **(f)** $\mu_\Delta = 0.2, \sigma_\Delta = 1$

These figures show the mean squared error (MSE) and the trace of theoretical asymptotic variance of $\hat{\theta}^\lambda$ as a function of the weight λ. The vertical gray line indicates the optimal weight λ^* that minimizes the theoretical asymptotic variance. In this simulation, we assume that the covariates follow a latent factor structure $X_{il} = \Lambda_i F_l + e_{il}$ and the outcomes $Y_i = X_i^\top \theta_0 + \epsilon_i$, where $F_l \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1), e_{il} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,4)$. The true coefficients are $\theta_{0l} = 1$ for all l. We set $N = 5000$, $N_0 = 500$, $L = 20$, $L_0 = 10$, and run 200 simulations for each setup.

$\theta_0^{(2)}\sigma_\Delta^4 + (\theta_0^{(1)})^2\mu_\Delta^2\sigma_\Delta^2\big)$.

Proposition 1 establishes that the estimator is asymptotically normal. When $r_1 > 0$, so that a positive fraction of units are partially observed, the asymptotic variance of the estimator depends on the weight λ. In this case, the optimal weight λ^* that minimizes the total asymptotic variance is

$$\lambda^* = \frac{\sigma_\epsilon^2 \cdot \operatorname{tr}(\Sigma_X^{-1})}{(\sigma_\epsilon^2 + \theta_0^{(2)}\sigma_\Delta^2 + (\theta_0^{(1)})^2\mu_\Delta^2)\operatorname{tr}(\Sigma_X^{-1}) + \operatorname{tr}(\Sigma_X^{-1}\Sigma_1\Sigma_X^{-1})}.$$

This expression shows that λ^* increases with the outcome noise σ_ϵ^2 and decreases with the imputation error, captured by $|\mu_\Delta|$ and σ_Δ. Intuitively, noisier outcomes make it more desirable to incorporate additional data, while larger imputation errors reduce the reliability of the imputed values. Notably, λ^* does not depend on the fraction r_1 of partially observed units. When $r_1 = 0$, the partially observed units are asymptotically negligible, and the asymptotic variance becomes independent of λ.

Figure 2 illustrates our theoretical results through a simulation. It shows the mean squared error (MSE) and the theoretical asymptotic variance of $\hat{\theta}^\lambda$ as functions of λ for different imputation qualities.

There are three main takeaways from Figure 2. First, we confirm that the optimal weight λ^* increases with the imputation quality (smaller $|\mu_\Delta|$ and σ_Δ). Second, the weight selection substantially affects the estimation efficiency. Using the optimal weight λ^* consistently outperforms the complete-case analysis ($\lambda = 0$). The efficiency gains increase as the quality of the imputation improves, but even with poor imputation ($\sigma_\Delta = 1$), incorporating the partially observed units with appropriate downweighting still yields better results. Third, our theoretically optimal λ^* values closely match the empirically optimal ones that minimize the MSEs, validating the practical applicability of our asymptotic theory.

3 Main Model and Estimation

We now generalize the framework of Section 2 to accommodate general moment-based estimation problems with heterogeneous missing covariate patterns. We formalize the general problem setup and notations in Section 3.1, and introduce our weighted debiased estimation approach in Section 3.2. Section 3.3 then describes two classes of imputation methods that can be incorporated within our framework, and Section 3.4 presents the cross-fitting procedure that ensures valid inference when imputation functions are learned from the data.

3.1 Problem Setup

We consider a general moment-based estimation problem. The target parameter $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ is defined as the unique solution to the population moment condition

$$G(\theta) = \mathbb{E}[g(Z_i; \theta)] = 0, \tag{2}$$

where $Z_i \coloneqq (Y_i, X_i)$ denotes the ith observation in an i.i.d. sample of outcome-covariate pairs. This formulation encompasses a broad class of estimation problems that can be expressed using the method of moments, including both linear and nonlinear models. More generally, the moment function $g(Z_i; \theta)$ may be overidentified, with $g(Z_i; \theta) \in \mathbb{R}^q$ for some $q \geq d$. For expositional clarity and notational simplicity, we focus on the just-identified case $q = d$ in the main framework. Section 4.4 shows that all results extend straightforwardly to the overidentified case within a standard GMM framework. For notational simplicity, we write $g_\theta(z) \coloneqq g(z; \theta)$.

Our primary focus is structural estimation, where the parameter θ_0 is the object of interest. In particular, we seek to understand how covariates affect outcomes, rather than to optimize predictive

accuracy. For example, in our empirical study we focus on the coefficient on carbon emissions, which captures the economic relationship between emissions and stock returns. Unlike predictive modeling, which can be assessed using held-out data, structural estimation generally does not have an analogous out-of-sample evaluation that directly validates the estimator. Consequently, we require estimation and inference that are valid for θ_0.

As in Section 2, the central challenge arises from missing covariates. While the outcomes Y_i are fully observed, the covariate vectors X_i contain missing entries. We denote the observation pattern for unit i by $W_i \in \{0,1\}^L$, where $W_{il} = 1$ indicates that covariate X_{il} is observed and $W_{il} = 0$ indicates that the covariate is missing. Across the data set, units may share identical observation patterns. We let $W^{(0)}, W^{(1)}, \ldots, W^{(m)} \in \{0,1\}^L$ denote the distinct unit-level observation patterns present in the sample, where $W^{(0)} = \mathbf{1}_L$ denotes the fully observed pattern and $W^{(1)}, \ldots, W^{(m)}$ denote the partially observed patterns. The number m is the number of distinct partially observed unit-level patterns in the data. For example, if missingness occurs in blocks as illustrated in Figure 1(a), then there is only one observation pattern $W^{(1)}$ shared by the partially observed units. In contrast, if units are missing covariates in heterogeneous ways, as in Figure 1(b), the number of distinct observation patterns m can be large.

For clarity of exposition, we begin with a setting where the observation pattern W_i is independent of the data Z_i. This allows us to present the core building blocks, debiasing and adaptive weighting, of the framework without additional distributional adjustments. In Section 5, we show that when missingness is endogenous but driven by observed covariates, the framework can be reformulated by reweighting the moment conditions to account for distributional differences across observation patterns. Additionally, we assume sufficiently many fully observed units for constructing the debiasing terms. Section 6 extends our framework to settings without fully observed units.

3.2 Weighted Debiased Estimation Framework

We now generalize the two key ideas from Section 2.3, data-driven debiasing and adaptive weighting, to accommodate general moment conditions and heterogeneous observation patterns. In this setting, heterogeneous observation patterns imply pattern-specific imputations, which affect both the debiasing terms and the weighting scheme.

For each partially observed pattern $W^{(j)}$, an imputation function f^j maps the covariate vector

and the pattern indicator to an imputed vector $\tilde{X}$:

$$f^j : (X, W^{(j)}) \mapsto \tilde{X}, \quad \text{where } \tilde{X}_l = \begin{cases} X_l, & \text{if } W_l^{(j)} = 1 \\ \hat{X}_l, & \text{if } W_l^{(j)} = 0 \end{cases},$$

and $\hat{X}_l$ denotes the imputed value for a missing covariate l, constructed using only the covariates that are observed under the corresponding pattern (those with $W_l^{(j)} = 1$). Crucially, each f^j depends only on observed covariates. As a result, f^j is well-defined for units with observation pattern $W^{(j)}$ as well as for units with additional observed covariates, including the fully observed units.

We apply the data-driven debiasing method and adaptive weighting scheme from Section 2.3 separately to each partially observed pattern. Let S_j denote the subset of units with observation pattern $W^{(j)}$, with sample size N_j, for $j = 0, 1, \ldots, m$, where S_0 corresponds to the fully observed subset. The resulting weighted debiased moment estimator is

$$\begin{aligned} G_N^{\lambda}(\theta) = \; & \frac{1}{N_0} \sum_{i \in S_0} g_\theta(X_i, Y_i) \\ & + \sum_{j=1}^{m} \lambda_j \frac{N_j}{N} \left(\frac{1}{N_j} \sum_{i \in S_j} g_\theta(f^j(X_i), Y_i) - \frac{1}{N_0} \sum_{i \in S_0} g_\theta(f^j(X_i), Y_i) \right). \end{aligned} \tag{3}$$

Estimator (3) closely parallels the illustrative example in Section 2. The first term uses only the fully observed units and provides an unbiased but inefficient baseline. Each subsequent term incorporates the imputed data from a partially observed subset S_j through the pattern-specific weight λ_j. The corresponding subtraction term, computed using the fully observed units, removes the systematic bias induced by imputation, thereby preserving the unbiasedness of the estimator.

The weights $\lambda = (\lambda_1, \ldots, \lambda_m)$ determine the contribution of each partially observed subset to the estimator and are chosen to improve statistical efficiency by accounting for heterogeneity in imputation reliability across patterns. Intuitively, imputation quality varies across observation patterns: patterns with more observed covariates typically yield more reliable imputations and can therefore receive larger weights, whereas patterns with more extensive missingness tend to produce noisier imputations and thus receive smaller weights. As shown in the illustrative example, setting all weights to zero recovers the complete-case estimator, while appropriately chosen positive weights can substantially improve estimation efficiency.

The estimator in (3) is valid for any finite number m of observation patterns. In practice, however,

the number m can be large, leading to computational and statistical challenges when estimating pattern-specific weights. To address this issue, we introduce a data-driven grouping approach that assigns a common weight to observation patterns with similar imputation quality. The details of this grouping procedure are provided in Appendix A.1. This dimension reduction preserves the key benefit of adaptive weighting while keeping the number of parameters tractable.

This grouping approach is fully compatible with the moment estimator in (3), where the pattern-specific weights are replaced by group-level weights without changing the debiasing structure. Formally, suppose that the m partially observed patterns are grouped into $\tilde{m}$ groups and assigned group-level weights $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{\tilde{m}}$. Let $\mathcal{I}_1, \ldots, \mathcal{I}_{\tilde{m}}$ denote the set of pattern indices in each group. The resulting grouped-weight estimator is

$$
\begin{aligned}
G_N^{\lambda}(\theta) = \; & \frac{1}{N_0} \sum_{i \in S_0} g_\theta(X_i, Y_i) \\
& + \sum_{h=1}^{\tilde{m}} \tilde{\lambda}_h \sum_{j \in \mathcal{I}_h} \frac{N_j}{N} \left(\frac{1}{N_j} \sum_{i \in S_j} g_\theta(f^j(X_i), Y_i) - \frac{1}{N_0} \sum_{i \in S_0} g_\theta(f^j(X_i), Y_i) \right).
\end{aligned} \tag{4}
$$

In practice, the number of groups $\tilde{m}$ can be chosen to be small (e.g., fewer than 10), reducing the number of weight parameters from m to $\tilde{m}$. For ease of exposition, all theoretical results in Section 4 are stated for the estimator with pattern-specific weights, under a fixed number of observation patterns m, which may be large but does not grow with the sample size. The grouping procedure is intended as a practical dimension-reduction method that improves numerical stability and finite-sample efficiency. Our simulation results suggest that a moderate number of weights is sufficient to achieve most efficiency gains in general settings (see Section 7).

The concept of using observed entries to quantify the imputation bias appears in the missing data literature, where synthetic missing patterns are generated on observed data to evaluate the imputation accuracy (Duan, Pelger, and Xiong, 2024a). However, our framework not only assesses the bias induced by imputation but also constructs explicit debiasing terms that enable valid inference even with imputed data. Our approach also closely relates to the prediction-powered inference (PPI) framework of Angelopoulos, Bates, Fannjiang, Jordan, and Zrnic (2023), which augments datasets with predicted outcomes. However, our setting introduces several novel challenges. First, we handle multivariate missingness in covariates rather than univariate outcome variables, which complicates both imputation and bias correction. Second, we accommodate heterogeneous observation patterns with pattern-specific (group-level) weights λ_j ($\tilde{\lambda}_j$), enabling us to adaptively incorporate partially

observed units based on their imputation quality. Third, our approach allows valid inference even without fully observed units (Section 6), extending beyond the scope of existing frameworks.

3.3 Imputation Methods

This section describes the imputation functions used in our framework and shows how different classes of imputation methods can be incorporated without affecting the validity of downstream inference. In practice, the imputation functions f^j in the weighted debiased estimator usually need to be learned from the data. Our framework accommodates a broad class of imputation methods that use observed covariates to infer missing ones. Here, we highlight two classes that are particularly relevant in practice. When auxiliary information beyond the observed covariates is available, it can be incorporated to further improve the imputation accuracy (see, e.g., Duan, Pelger, and Xiong (2024b)). However, for simplicity of notation, we present the methods using only the covariates X_i.

Regression-based imputation. Regression-based imputation learns the dependence structure among covariates through supervised regression models trained on observed entries. Specifically, for each observation pattern $W^{(j)}$, we train an imputation function f^j to minimize $\mathbb{E}[\|f^j(X_i, W^{(j)}) - X_i\|^2]$ using fully observed units. The function form of f^j is flexible and can range from simple linear regressions to nonlinear machine learning models such as deep neural networks or random forests. These methods are easy to implement and often perform well when sufficient fully observed data are available.

Factor-based imputation. A complementary and widely used class of imputation methods exploits the low-rank structure in high-dimensional covariate panels, as commonly observed in economic and financial data (Mazumder, Hastie, and Tibshirani, 2010; Candès and Recht, 2012). By modeling the covariate panel using an approximate latent factor structure, factor-based imputation captures dependence across both units and covariates (Bai and Ng, 2021; Jin, Miao, and Su, 2021; Xiong and Pelger, 2023). This approach has demonstrated strong empirical performance in economics and finance applications (Bryzgalova, Lerner, Lettau, and Pelger, 2025). Importantly, factor-based imputation remains applicable even when no units are fully observed, a setting where regression-based methods typically fail and which we formally address in Section 6.

To provide a concrete example of this class of methods, we consider a factor-model imputation approach, which we also adopt in our empirical study. The factor-model imputation approach assumes

that the covariate panel $\mathbf{X} = [X_1, \ldots, X_N]^\top \in \mathbb{R}^{N \times L}$ can be well approximated by a latent factor structure with k common factors:

$$X_{il} = \Lambda_i^\top F_l + \epsilon_{il}, \quad i = 1, \ldots, N,\ l = 1, \ldots, L,$$

where $F_l \in \mathbb{R}^k$ is a k-dimensional vector of latent factors for covariate l, $\Lambda_i \in \mathbb{R}^k$ is the loading for unit i, and ϵ_{il} is unit i's idiosyncratic error for the l-th covariate. We treat the factors as fixed parameters and assume that the number of factors k is known (it can be estimated in practice). The loadings Λ_i and idiosyncratic errors $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iL})$ are drawn independently from distributions P_Λ and P_ϵ, respectively, ensuring that the covariate vectors are i.i.d. across units.

Following Xiong and Pelger (2023), we estimate the latent factor structure using the partially observed covariate panel. Specifically, we first construct a cross-covariate second-moment matrix $\hat{\Sigma}$ using pairs of jointly observed covariates. Let $Q_{lh} = \{i : W_{il} = W_{ih} = 1\}$ denote the set of units where both covariates l and h are observed. The (l, h)-th entry of the estimated matrix $\hat{\Sigma}$ is

$$\hat{\Sigma}_{lh} = \frac{1}{|Q_{lh}|} \sum_{i \in Q_{lh}} X_{il} X_{ih}.$$

We then estimate the factors $\hat{F}$ through principal component analysis (PCA) of $\hat{\Sigma}$. Under the standard identification assumption that $\hat{F}^\top \hat{F}/L = I_k$, the estimated factors are given by $\sqrt{L}$ times the eigenvectors corresponding to the k largest eigenvalues of $\hat{\Sigma}/L$. The imputation function $f^j(\cdot)$ imputes missing covariates by projecting the observed covariates onto the estimated factor space. Specifically,

$$f_l^j(X, W^{(j)}) = \begin{cases} X_l, & W_l^{(j)} = 1 \\ \hat{F}_l^\top \left(\sum_{h=1}^L W_h^{(j)} \hat{F}_h \hat{F}_h^\top \right)^{-1} \left(\sum_{h=1}^L W_h^{(j)} \hat{F}_h X_h \right), & W_l^{(j)} = 0 \end{cases}.$$

3.4 Cross-Fitting

This section introduces the cross-fitting procedure that ensures unbiasedness of the weighted debiased estimator when the imputation functions are estimated from the data. For the estimator in equation (3) to be unbiased, the imputation functions f^j need to be independent of the data on which they are evaluated. This requirement arises because we adopt a data-driven debiasing approach that uses observed covariates to estimate and remove the bias induced by imputation.

Specifically, in (3), unbiasedness requires that

$$\mathbb{E}[g_\theta(f^j(X_i), Y_i) \mid i \in S_j] = \mathbb{E}[g_\theta(f^j(X_i), Y_i) \mid i \in S_0].$$

Under the assumption that units are i.i.d. across observation patterns, this equality holds provided that the imputation functions are trained independently of the units used for evaluation. While closed-form debiasing approaches can avoid this independence requirement, they typically rely on parametric assumptions on the data-generating process and are, therefore, limited to specific imputation models. In contrast, our data-driven debiasing approach does not impose such assumptions and can accommodate general imputation methods.

We achieve this independence through K-fold cross-fitting. Specifically, we partition the N units into K equal-sized folds $\tilde{S}_1, \ldots, \tilde{S}_K$, ensuring that each fold contains the same proportion of units from each of the $m+1$ observation patterns.[6] Let $S_{j,k} \coloneqq S_j \cap \tilde{S}_k$ denote the set of units with pattern j in fold k. For each fold k, we train imputation functions

$$f^j_{-k}(\cdot) \coloneqq f^j(\cdot; [N] \backslash \tilde{S}_k)$$

using data from all folds except $\tilde{S}_k$ and apply them to the units in the held-out fold $\tilde{S}_k$. Aggregating estimates across all folds yields the weighted debiased moment estimator

$$\begin{aligned} G^\lambda_N(\theta) = \; & \frac{1}{N_0} \sum_{i \in S_0} g_\theta(X_i, Y_i) \\ & + \sum_{k=1}^{K} \sum_{j=1}^{m} \lambda_j \frac{N_j}{N} \left(\frac{1}{N_j} \sum_{i \in S_{j,k}} g_\theta(f^j_{-k}(X_i), Y_i) - \frac{1}{N_0} \sum_{i \in S_{0,k}} g_\theta(f^j_{-k}(X_i), Y_i) \right). \end{aligned} \tag{5}$$

By construction, each imputation function is evaluated only on observations that were not used in its training. As a result, even though the imputation functions are estimated from the data, they remain independent of the observations used in moment estimation.

In the remainder of the paper, we refer to the *weighted debiased moment estimator* $G^\lambda_N(\theta)$ as the estimator defined in (5), and denote by $\hat{\theta}^\lambda$ the solution to $G^\lambda_N(\theta) = 0$. In Section 4, we establish the asymptotic normality of this estimator under general regularity conditions.

Our framework also extends naturally to panel data settings, where each unit is observed over time. In such cases, the imputation function may incorporate temporal information, such as lagged

[6] Without loss of generality, we assume N_j/K is an integer for each subset S_j.

covariates, to improve imputation accuracy. The weighted debiased moment construction and the asymptotic results established in Section 4 extend to this setting with only minor modifications. The details are provided in Appendix A.2.

4 Theoretical Results

This section establishes the theoretical properties of the proposed weighted debiased estimator. We first derive consistency and asymptotic normality for the estimator with general weights in Section 4.1. We then derive the optimal weights that minimize the asymptotic variance in Section 4.2 and examine the stability conditions required for our inferential theory in Section 4.3. Finally, Section 4.4 extends the framework to overidentified moment conditions within a generalized method of moments (GMM) setting.

4.1 Inferential Theory

We begin by developing the inferential theory for the weighted debiased estimator $\hat{\theta}^{\lambda}$. For notational simplicity and clarity of exposition, we impose an exogenous missingness assumption in this section. Section 5 shows how to accommodate endogenous missingness that depends on observed covariates by reweighting the moment conditions.

Assumption 1 (Exogenous missingness)**.** *The observation pattern W is independent of the data Z.*

To establish the asymptotic properties of $\hat{\theta}^{\lambda}$, we impose standard regularity conditions on the moment function g_θ evaluated on both the observed data and the imputed covariates, as stated in Assumption 2.

Assumption 2 (Moment function regularity)**.**

1. *(Finite variance). For any $\theta \in \Theta$, $\mathrm{Var}(g_\theta(X_i, Y_i))$ and $\mathrm{Var}\big(g_\theta(f^j_{-k}(X_i), Y_i)\big)$ are finite for all j and any given imputation function f^j_{-k}.*
2. *(Differentiability). The functions $\theta \mapsto g_\theta(x, y)$ and $\theta \mapsto g_\theta(f^j_{-k}(x), y)$ are continuous for all (x, y) and functions f^j_{-k}. Moreover, $G(\theta) = \mathbb{E}[g_\theta(X_i, Y_i)]$ is differentiable at θ_0 with positive definite Jacobian $J_{\theta_0} = \nabla G_\theta(\theta_0)$.*
3. *(Local Lipschitz continuity). There exists a neighborhood of θ_0 such that for all $\theta_1, \theta_2 \in \Theta$ in*

this neighborhood,

$$\|g_{\theta_1}(x,y) - g_{\theta_2}(x,y)\| \le L(x,y)\cdot\|\theta_1-\theta_2\|,$$
$$\left\|g_{\theta_1}(f^j_{-k}(x),y) - g_{\theta_2}(f^j_{-k}(x),y)\right\| \le L_j(x,y)\cdot\|\theta_1-\theta_2\|,$$

where $\mathbb{E}\|L(X_i,Y_i)\|^2 < \infty$ *and* $\mathbb{E}\|L_j(X_i,Y_i)\|^2 < \infty$ *for all* j.

Assumption 2 collects standard conditions for moment-based estimation. Assumption 2.1 requires the moment functions to have finite variance, ensuring the applicability of central limit theorems. Assumption 2.2 guarantees local identification. Assumption 2.3 imposes local Lipschitz continuity of the moment functions, which enables linearization of the empirical moment process around θ_0.

Theorem 1 (Consistency)**.** *Suppose* $N_0 \to \infty$ *as* $N \to \infty$*, and Assumptions 1 and 2 hold. For given weights* $\lambda = (\lambda_1, \ldots, \lambda_m)$*, assume* $\hat{\theta}^\lambda$ *uniquely solves* $G_N^\lambda(\theta) = 0$ *with probability tending to one. In addition, assume that either (1) the parameter space* Θ *is compact and Assumption 2.3 holds globally on* Θ*, or (2)* $G_N^\lambda(\theta)$ *is nondecreasing, i.e.,* $(G_N^\lambda(\theta_1) - G_N^\lambda(\theta_2))^\top(\theta_1-\theta_2) \ge 0$ *for any* $\theta_1, \theta_2 \in \Theta$*. Then,* $\hat{\theta}^\lambda$ *is consistent, i.e.,* $\hat{\theta}^\lambda \xrightarrow{p} \theta_0$ *as* $N \to \infty$.

Theorem 1 establishes the consistency of the weighted debiased estimator for any fixed weights λ_j. Building on this result, we derive its asymptotic distribution.

Theorem 2 (Asymptotic normality)**.** *Let* $r_j = \lim_{N\to\infty} N_j/N$*. For each* k *and* j *with* $N_j \to \infty$*, assume there exists a function* $\bar{f}^j$ *satisfying*

$$\begin{aligned}\frac{1}{\sqrt{N_j}}\sum_{i\in S_{j,k}} \big(&g_{\theta_0}(f^j_{-k}(X_i),Y_i) - \mathbb{E}_Z[g_{\theta_0}(f^j_{-k}(X_i),Y_i)] \\ &- \big(g_{\theta_0}(\bar{f}^j(X_i),Y_i) - \mathbb{E}[g_{\theta_0}(\bar{f}^j(X_i),Y_i)]\big)\big) = o_p(1).\end{aligned} \tag{6}$$

Under the conditions of Theorem 1 and condition (6), $\hat{\theta}^\lambda$ *is asymptotically normal:*

$$\sqrt{N_0}(\hat{\theta}^\lambda - \theta_0) \xrightarrow{d} \mathcal{N}\Bigg(0, \underbrace{J_{\theta_0}^{-1}\left(\Sigma^\lambda_{\Delta,0} + \sum_{j=1}^m \lambda_j^2 r_j r_0 \Sigma_{f,j}\right) J_{\theta_0}^{-1}}_{:=\Sigma^\lambda}\Bigg), \tag{7}$$

with $\Sigma^\lambda_{\Delta,0} = \mathrm{Cov}\left(g_{\theta_0}(X_i,Y_i) - \sum_{j=1}^m \lambda_j r_j g_{\theta_0}(\bar{f}^j(X_i),Y_i)\right)$ *and* $\Sigma_{f,j} = \mathrm{Cov}\left(g_{\theta_0}(\bar{f}^j(X_i),Y_i)\right)$.

Theorem 2 establishes the asymptotic normality of the estimator, which provides the foundation for constructing valid confidence intervals and conducting hypothesis testing. The convergence rate

$\sqrt{N_0}$ is driven by the number of fully observed units, which are used to construct the debiasing terms that correct for imputation errors in the partially observed samples.

The asymptotic variance Σ^λ consists of two components. The first term $\Sigma^\lambda_{\Delta,0}$, which is the covariance of the difference between the complete-case and weighted imputed moments, captures variance reduction from incorporating imputed observations. The second term, $\sum_{j=1}^m \lambda_j^2 r_j r_0 \Sigma_{f,j}$, quantifies additional variance introduced by these observations. The weights λ_j govern this efficiency trade-off by determining how much information from each partially observed group is incorporated into the estimator. The influence of a partially observed subset j on the asymptotic distribution depends on its relative size r_j: when $N_j/N \to r_j > 0$, subset j contributes to the estimator for $\lambda_j > 0$, whereas when $r_j = 0$, its contribution vanishes asymptotically regardless of the weight λ_j. Section 4.2 derives the optimal weights λ^* that minimize this asymptotic variance.

The additional condition (6) in Theorem 2 ensures that the randomness in the imputation functions introduced by training samples is asymptotically negligible, enabling valid inference under the cross-fitting procedure. Section 4.3 analyzes the conditions under which this requirement holds.

4.2 Optimal Weights

This subsection characterizes the choice of weight vector λ that optimally trades off the efficiency gains from incorporating imputed observations against their additional variance. As shown in Theorem 2, the weights directly enter the asymptotic variance of the weighted debiased estimator and play a central role in determining its statistical efficiency.

In principle, the weights can be chosen to optimize different criteria derived from the asymptotic covariance matrix, such as the total variance, the variance of a subset of parameters of interest, or other criteria including eigenvalue-based measures. In this paper, we focus on minimizing the total asymptotic variance, which yields a transparent quadratic objective and admits a closed-form characterization of the optimal weights. For notational convenience, define $g_\theta^{(j)} \coloneqq g_{\theta_0}(\bar{f}^j(X_i), Y_i)$. The next corollary formalizes the resulting optimization problem and its solution.

Corollary 1 (Optimal weights)**.** *Under the conditions of Theorem 2, the total asymptotic variance* $\mathrm{tr}(\Sigma^\lambda)$ *is a quadratic function of* $\lambda = (\lambda_1, \ldots, \lambda_m)$ *and can be written as*

$$\mathrm{tr}(\Sigma^\lambda) = \lambda^\top Q \lambda - 2v^\top \lambda + \mathrm{tr}(J_{\theta_0}^{-1} \mathrm{Cov}(g_{\theta_0}) J_{\theta_0}^{-1}),$$

where $v \in \mathbb{R}^m$ *has components* $v_j = \frac{1}{2} r_j \cdot \mathrm{tr}(J_{\theta_0}^{-1}(\mathrm{Cov}(g_{\theta_0}, g_{\theta_0}^{(j)}) + \mathrm{Cov}(g_{\theta_0}^{(j)}, g_{\theta_0}))J_{\theta_0}^{-1})$ *and* $Q \in \mathbb{R}^{m \times m}$

is positive semidefinite with entries

$$Q_{jh} = \begin{cases} r_j(r_0 + r_j) \cdot \mathrm{tr}\left(J_{\theta_0}^{-1}\mathrm{Cov}(g_{\theta_0}^{(j)})J_{\theta_0}^{-1}\right), & j = h \\ r_j r_h \cdot \mathrm{tr}\left(J_{\theta_0}^{-1}\mathrm{Cov}(g_{\theta_0}^{(j)}, g_{\theta_0}^{(h)})J_{\theta_0}^{-1}\right), & j \neq h. \end{cases}$$

The quadratic objective is minimized over $\lambda \in \mathbb{R}^m$ *by* $\lambda^* = Q^+ v$*, where* Q^+ *denotes the pseudoinverse of* Q*. The minimized total asymptotic variance is* $\mathrm{tr}(\Sigma^{\lambda^*}) = \mathrm{tr}(J_{\theta_0}^{-1}\mathrm{Cov}(g_{\theta_0})J_{\theta_0}^{-1}) - \lambda^{*\top} Q \lambda^*$.

Remark 1. *When a common weight* $\bar{\lambda}$ *is imposed across all partially observed subsets, i.e.,* $\lambda_1 = \cdots = \lambda_m = \bar{\lambda}$*, the optimal scalar weight simplifies to*

$$\bar{\lambda}^* = \frac{\sum_{j=1}^m r_j \mathrm{tr}\left(J_{\theta_0}^{-1}\left(\mathrm{Cov}(g_{\theta_0}, g_{\theta_0}^{(j)}) + \mathrm{Cov}(g_{\theta_0}^{(j)}, g_{\theta_0})\right)J_{\theta_0}^{-1}\right)}{2\left(\sum_{j,h=1}^m r_j r_h \mathrm{tr}\left(J_{\theta_0}^{-1}\mathrm{Cov}(g_{\theta_0}^{(j)}, g_{\theta_0}^{(h)})J_{\theta_0}^{-1}\right) + \sum_{j=1}^m r_j r_0 \mathrm{tr}\left(J_{\theta_0}^{-1}\mathrm{Cov}(g_{\theta_0}^{(j)})J_{\theta_0}^{-1}\right)\right)}.$$

Corollary 1 provides two key insights for weight selection. First, the optimal weights λ_j^* depend on both the quality of imputation, captured by the covariance between the complete-case moment g_{θ_0} and the imputed moment $g_{\theta_0}^{(j)}$, and the covariance structure among the imputed moments themselves. Higher-quality imputations lead to larger optimal weights, while greater variability induced by imputation reduces the weight placed on the corresponding subset. In the special case of perfect imputation, where $g_{\theta_0}^{(j)} = g_{\theta_0}$ for all j, the optimal weights satisfy $\lambda_j^* = 1$, and the estimator recovers the oracle estimator based on fully observed data. Conversely, when the imputed moments are weakly informative and nearly uncorrelated with the complete-case moment, the optimal weights shrink toward zero, and the estimator reduces to the complete-case analysis.

Second, this optimal weighting scheme guarantees an efficiency improvement over the complete-case analysis. Specifically, we have $\mathrm{tr}(\Sigma^{\lambda^*}) \leq \mathrm{tr}(J_{\theta_0}^{-1}\mathrm{Cov}(g_{\theta_0})J_{\theta_0}^{-1})$, with strict inequality when the partially observed subsets provide informative imputed moments. Even under the common-weight restriction in Remark 1, incorporating imputed data improves estimation efficiency. However, as we demonstrate through simulations in Section 7, allowing heterogeneous weights can deliver substantially larger efficiency gains.

In applications, the optimal weights λ^* are unknown and must therefore be estimated. We construct a data-driven estimator $\hat{\lambda}$ and define $\hat{\theta}^{\hat{\lambda}}$ as the solution to $G_N^{\hat{\lambda}}(\theta) = 0$. We consider two approaches for estimating λ^*. The first approach directly estimates Q and v in Corollary 1 and sets $\hat{\lambda} = \hat{Q}^+ \hat{v}$. The second approach estimates the asymptotic covariance matrix in Theorem 2 and chooses $\hat{\lambda}$ to optimize a user-specified objective, allowing criteria beyond total variance minimization. When

additional constraints are imposed, such as $\lambda_j \in [0,1]$, we compute $\hat{\lambda}$ by solving the corresponding constrained quadratic program. The following corollary shows that using a consistent data-driven weight $\hat{\lambda}$ does not affect the asymptotic distribution of the resulting estimator.

Corollary 2 (Inference with estimated weights)**.** *Let $\hat{\lambda}$ be a random weight vector taking values in a compact set $\Lambda \subset \mathbb{R}^m$ such that $\hat{\lambda} \xrightarrow{p} \lambda^*$ for some deterministic $\lambda^* \in \Lambda$. Under the conditions of Theorem 2, $\sqrt{N_0}\,(\hat{\theta}^{\hat{\lambda}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma^{\lambda^*})$, where Σ^{λ^*} is the asymptotic covariance matrix evaluated at λ^*.*

Corollary 2 shows that using a consistent estimator $\hat{\lambda}$ does not affect the limiting distribution of the estimator. For the plug-in approach $\hat{\lambda} = \hat{Q}^{+}\hat{v}$, $\hat{\lambda}$ is consistent provided that $\hat{v} \xrightarrow{p} v$, $\hat{Q} \xrightarrow{p} Q$, the smallest nonzero eigenvalue of Q is bounded away from zero, and $\operatorname{rank}(\hat{Q}) = \operatorname{rank}(Q)$ with probability tending to one. Alternatively, for the second approach, consistency holds if the estimated objective converges uniformly to a deterministic limit with a unique minimizer over the set Λ.

We now describe the estimation of several key components required for implementation. The Jacobian J_{θ_0} is estimated using an initial estimator $\hat{\theta}_{\text{init}}$ (e.g., the complete-case estimator) as

$$\hat{J} = \frac{1}{N_0} \sum_{i \in S_0} \nabla g_{\hat{\theta}_{\text{init}}}(X_i, Y_i).$$

To estimate the covariance matrices $\operatorname{Cov}(g_{\theta_0}^{(j)}, g_{\theta_0}^{(h)})$, we use a bootstrap procedure designed to capture both sampling variability and the randomness induced by training the imputation functions. Specifically, in each bootstrap iteration $b = 1, \ldots, B$, we randomly draw N_j/K units from each subset S_j to form a bootstrap sample $S^{(b)}$, train the imputation functions $f^j_{-S^{(b)}}(\cdot) = f^j(\cdot; [N] \backslash S^{(b)})$ on the complementary data, and apply them to the held-out units. The resulting covariance estimator is

$$\widehat{\operatorname{Cov}}(g_{\theta_0}^{(j)}, g_{\theta_0}^{(h)}) = \frac{1}{\sum_{b=1}^{B} |S_0^{(b)}|} \sum_{b=1}^{B} \sum_{i \in S_0^{(b)}} \left(g_{\hat{\theta}_{\text{init}}}(f^j_{-S^{(b)}}(X_i), Y_i) - \bar{g}^j_{\hat{\theta}_{\text{init}}} \right) \left(g_{\hat{\theta}_{\text{init}}}(f^h_{-S^{(b)}}(X_i), Y_i) - \bar{g}^h_{\hat{\theta}_{\text{init}}} \right)^{\top},$$

where $S_0^{(b)}$ denotes the set of fully observed units in $S^{(b)}$ and $\bar{g}^j_{\hat{\theta}_{\text{init}}}$ denotes the empirical mean of $g_{\hat{\theta}_{\text{init}}}(f^j_{-S^{(b)}}(X_i), Y_i)$.

4.3 Stability Condition

We now examine the stability condition (6) required to establish the asymptotic normality result in Theorem 2. This condition plays a key role in ensuring that the randomness introduced by estimating the imputation functions does not affect the limiting distribution of the estimator. In particular, condition (6) guarantees that the randomness of the cross-fitted imputation functions f^j_{-k}, arising from their dependence on the training sample $[N]\backslash\tilde{S}_k$, is asymptotically negligible. As a result, each random imputation function can be approximated by a deterministic function $\bar{f}^j$ in large samples, which is crucial for establishing the asymptotic normality of the cross-fitted estimator.

To analyze when this condition holds, we decompose each imputation function into a deterministic component and a stochastic remainder:

$$f^j_{-k}(x) = f^j(x; [N]\backslash\tilde{S}_k) = h^j(x) + \Delta^j(x; [N]\backslash\tilde{S}_k), \tag{8}$$

where $h^j(\cdot)$ denotes a deterministic function and $\Delta^j_{-k}(\cdot) = \Delta^j(\cdot; [N]\backslash\tilde{S}_k)$ represents the random fluctuation induced by sampling variability in the training data. The following proposition provides sufficient conditions on the magnitude of the stochastic component Δ^j_{-k} under which the stability condition (6) holds.

Proposition 2 (Sufficient conditions for (6))**.** *Assume the moment function g_{θ_0} satisfies the mean continuity condition: there exist $q \geq 1$ and a constant $C < \infty$ such that*

$$\left(\mathbb{E}\left[\|g_{\theta_0}(X_1, Y) - g_{\theta_0}(X_2, Y)\|^2\right]\right)^{1/2} \leq C \cdot \left(\mathbb{E}[\|X_1 - X_2\|^q]\right)^{1/q}. \tag{9}$$

For any pattern j and fold k, suppose the learned imputation function admits the decomposition (8) and satisfies

$$\mathbb{E}\left[\|\Delta^j_{-k}(X_i)\|^q\right] = o(1) \tag{10}$$

as $N \to \infty$. Then the stability condition (6) holds.

Proposition 2 states that a sufficient condition is that the stochastic component of the imputation function induced by the training sample to be asymptotically negligible as the sample size grows. Importantly, this does not impose consistency or asymptotic normality of the imputed values. If an imputation method involves internal randomness (e.g., random forests), one may either fix the random seed and verify stability with respect to sampling variability, or increase the ensemble size

so that the algorithmic randomness becomes negligible. The mean continuity condition (9) controls how perturbations in the covariates propagate into the moment function. This assumption is mild and holds for a broad class of moment functions used in practice. For example, in linear regression moments, condition (9) is satisfied with $q = 4$ under standard moment assumptions.

To demonstrate the practical relevance of Proposition 2, we show below that the factor model imputation method introduced in Section 3.3 satisfies condition (10) under an approximate latent factor structure, which in turn implies condition (6).

Proposition 3 (Stability of factor model imputation)**.** *Suppose the covariate panel* $\mathbf{X}$ *follows the approximate latent factor model*

$$X_{il} = \Lambda_i F_l + e_{il}, \quad i = 1, \ldots, N,\ l = 1, \ldots, L,$$

where loadings $\Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\Lambda^2)$*, idiosyncratic errors* $e_{il} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$*, and the factors* F_l *satisfy the normalization* $F^\top F/L = 1$ *with* $\min_l |F_l| \geq c > 0$*. Suppose Assumption 1 holds and there exists* $c > 0$ *such that* $|Q_{hl}|/N \geq c$ *for all* h, l*, where* $Q_{hl} = \{i : W_{ih} = W_{il} = 1\}$*. Then the factor model imputation method described in Section 3.3 satisfies condition* (10)*.*

Proposition 3 establishes that the factor model imputation method satisfies the stability requirement in Proposition 2. More generally, this stability condition can be verified for a broad class of commonly used regression-based imputation methods under standard regularity assumptions. For example, ℓ_2-regularized regression methods satisfy algorithmic stability under strong convexity and Lipschitz loss functions; tree-based methods satisfy stability under structural regularity conditions such as honesty, appropriate subsampling rates, and controlled leaf sizes; and high-dimensional sparse regression methods satisfy analogous concentration results under sparsity and restricted eigenvalue conditions (e.g., Bousquet and Elisseeff (2002); Györfi, Kohler, Krzyżak, and Walk (2002); Bühlmann and van de Geer (2011); Wager and Athey (2018)).

4.4 Overidentified GMM

This subsection extends the proposed framework to settings with overidentified moment conditions, where the moment function $g(Z_i; \theta)$ is q-dimensional with $q > d$.

In this setting, we continue to construct the weighted debiased moment estimator $G_N^\lambda(\theta) \in \mathbb{R}^q$ as defined in Equation (5), using the subset weights $\lambda = (\lambda_1, \ldots, \lambda_m)$. However, instead of defining

$\hat{\theta}^{\lambda}$ as the solution to $G_N^{\lambda}(\theta) = 0$, we estimate θ_0 using a generalized method of moments (GMM) criterion. Specifically, we define the estimator $\hat{\theta}_A^{\lambda}$ as

$$\hat{\theta}_A^{\lambda} = \arg\min_{\theta \in \Theta} G_N^{\lambda}(\theta)^{\top} A G_N^{\lambda}(\theta),$$

where A is a positive definite weighting matrix.

Under the conditions of Theorem 2, the weighted debiased moment admits a central limit theorem:

$$\sqrt{N_0} \cdot G_N^{\lambda}(\theta_0) \xrightarrow{d} \mathcal{N}(0, \Omega_{\lambda}),$$

where the asymptotic covariance $\Omega_{\lambda} \in \mathbb{R}^{q \times q}$ is given by

$$\Omega_{\lambda} = \Sigma_{\Delta,0}^{\lambda} + \sum_{j=1}^{m} \lambda_j^2 r_j r_0 \Sigma_{f,j},$$

where $\Sigma_{\Delta,0}^{\lambda} = \mathrm{Cov}(g_{\theta_0}(X_i, Y_i) - \sum_{j=1}^{m} \lambda_j r_j \, g_{\theta_0}(\bar{f}_j(X_i), Y_i))$ and $\Sigma_{f,j} = \mathrm{Cov}(g_{\theta_0}(\bar{f}_j(X_i), Y_i))$. Let $J_{\theta_0} = \nabla_{\theta} G(\theta_0) \in \mathbb{R}^{q \times d}$ denote the Jacobian of the population moments and assume that J_{θ_0} has full column rank.

By standard GMM arguments, the minimizer $\hat{\theta}_A^{\lambda}$ is consistent for θ_0 and satisfies

$$\sqrt{N_0}\,(\hat{\theta}_A^{\lambda} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V_{\lambda,A}),$$

where the asymptotic covariance matrix takes the form

$$V_{\lambda,A} = \left(J_{\theta_0}^{\top} A J_{\theta_0}\right)^{-1} J_{\theta_0}^{\top} A \, \Omega_{\lambda} \, A J_{\theta_0} \left(J_{\theta_0}^{\top} A J_{\theta_0}\right)^{-1}.$$

Within the class of estimators that combine observed and imputed data through the weighted debiased moment function $G_N^{\lambda}(\theta)$, the asymptotically efficient choice of the weighting matrix is $A^* = \Omega_{\lambda}^{-1}$. Under this choice, the asymptotic covariance simplifies to

$$V_{\lambda,A^*} = \left(J_{\theta_0}^{\top} \Omega_{\lambda}^{-1} J_{\theta_0}\right)^{-1}.$$

We can therefore select the weight vector λ by treating V_{λ,A^*} as the relevant asymptotic covariance matrix and optimizing a user-specified criterion, as described in Section 4.2. In implementation, we replace A by $\hat{A} = \hat{\Omega}_{\hat{\lambda}}^{-1}$. When $\hat{\lambda}$ converges in probability to λ^* and $\hat{\Omega}_{\hat{\lambda}}$ is a consistent estimator of

Ω_{λ^*}, then standard arguments imply

$$\sqrt{N_0}\,(\hat{\theta}^{\lambda}_{\hat{A}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V_{\lambda,A^*})\,.$$

5 Endogenous Missingness

In many empirical applications, the observation pattern of covariates may depend on the underlying data. For example, smaller firms may lack the resources to collect and report specific covariates compared to larger firms, resulting in systematic differences across observation patterns.

In our framework, we consider the endogenous missingness driven by fully observed covariates. Let O_i denote such covariates and assume that

$$W_i \perp\!\!\!\perp (Y_i, X_i) \mid O_i.$$

Our framework accommodates this form of endogenous missingness by reweighting the moment conditions using likelihood ratios. In particular, the only modification relative to Sections 3 and 4 is this weight adjustment; all arguments and derivations proceed analogously under the resulting reweighted measure.

Under this conditional independence structure, the joint distribution associated with each observation pattern can be decomposed as a pattern-specific marginal distribution of O_i, denoted as P^j_O, and a common conditional distribution of (Y_i, X_i) given O_i. As a result, distributional differences across observation patterns can be corrected by reweighting using the likelihood ratios of O_i. Incorporating these likelihood ratios into the estimator defined in equation (5) yields the following reweighted debiased moment estimator:

$$\begin{aligned} G^{\lambda}_N(\theta) = \; & \frac{1}{N_0}\sum_{i\in S_0} g_\theta(X_i, Y_i) \sum_{j=0}^{m} \frac{N_j}{N}\omega^j(O_i) \\ & + \sum_{k=1}^{K}\sum_{j=1}^{m} \lambda_j \frac{N_j}{N}\left(\frac{1}{N_j}\sum_{i\in S_{j,k}} g_\theta(f^j_{-k}(X_i), Y_i) - \frac{1}{N_0}\sum_{i\in S_{0,k}} g_\theta(f^j_{-k}(X_i), Y_i)\omega^j(O_i)\right), \end{aligned} \tag{11}$$

where the likelihood ratios $\omega^j(o) \coloneqq dP^j_O/dP^0_O(o)$ for $j = 1, \ldots, m$. In the first term, $\sum_{j=0}^{m}(N_j/N)\omega^j(O_i)$ reweights the fully observed units to represent the distribution across observation patterns, correcting for potential selection bias in the fully observed subset. In the remaining terms, the likelihood ratios adjust the bias correction terms to account for distributional differences across patterns.

The following theorem establishes the asymptotic properties of the reweighted estimator.

Theorem 3 (Asymptotic theory for reweighted estimator (11))**.** *Suppose $N_0 \to \infty$ as $N \to \infty$, $W_i \perp\!\!\!\perp (Y_i, X_i) \mid O_i$ and Assumption 2 holds. For given weights $\lambda = (\lambda_1, \ldots, \lambda_m)$, assume $\hat{\theta}^\lambda$ uniquely solves $G_N^\lambda(\theta) = 0$ in (11) with probability tending to one. In addition, assume that dP_O^j is absolutely continuous with respect to dP_O^0 and $\sup_{o \in \mathcal{O}} \omega^j(o) \leq M < \infty$. Assume that either (1) the parameter space Θ is compact and Assumption 2.3 holds globally on Θ, or (2) $G_N^\lambda(\theta)$ is nondecreasing, i.e., $(G_N^\lambda(\theta_1) - G_N^\lambda(\theta_2))^\top(\theta_1 - \theta_2) \geq 0$ for any $\theta_1, \theta_2 \in \Theta$. Then, $\hat{\theta}^\lambda$ is consistent, i.e., $\hat{\theta}^\lambda \xrightarrow{p} \theta_0$ as $N \to \infty$. Let $r_j = \lim_{N\to\infty} N_j/N$. Moreover, assume the stability condition (6) holds for both $g_{\theta_0}(f_{-k}^j(X_i), Y_i)$ and $g_{\theta_0}(f_{-k}^j(X_i), Y_i)\omega^j(O_i)$. Then, $\hat{\theta}^\lambda$ is asymptotically normal:*

$$\sqrt{N_0}(\hat{\theta}^\lambda - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, J_{\theta_0}^{-1}\left(\Sigma_{\Delta,0}^\lambda + \sum_{j=1}^m \lambda_j^2 r_j r_0 \Sigma_{f,j}\right) J_{\theta_0}^{-1}\right), \tag{12}$$

where the matrices $\Sigma_{\Delta,0}^\lambda = \mathrm{Cov}_{P^0}\left(\sum_{j=0}^m r_j g_{\theta_0}(X_i, Y_i)\omega^j(O_i) - \sum_{j=1}^m \lambda_j r_j g_{\theta_0}(\bar{f}^j(X_i), Y_i)\omega^j(O_i)\right)$ and $\Sigma_{f,j} = \mathrm{Cov}_{P^j}(g_{\theta_0}(\bar{f}^j(X_i), Y_i))$.

Theorem 3 extends the main inferential results to settings with endogenous missingness driven by fully observed covariates. The proof follows arguments analogous to those used in Theorems 1 and 2. The boundedness of the likelihood ratios ensures sufficient overlap across distributions, preventing instability caused by extreme weights. The stability condition guarantees that the randomness induced by estimating the imputation functions remains asymptotically negligible. Theorem 3 is stated for oracle likelihood ratios. When these ratios are estimated, similar results hold under additional conditions on the convergence rate of the likelihood ratio estimators.

While reweighting eliminates bias arising from endogenous missingness, it also introduces additional sources of variation and estimation error. First, estimating the likelihood ratios $\omega^j(o)$ introduces additional estimation error. When the dimension of O_i is small, kernel density estimation can be used to directly estimate the density ratios. In higher-dimensional settings, where density estimation becomes infeasible, we leverage the relationship

$$\omega^j(o) = \frac{\mathbb{P}(W_i = W^{(j)} \mid O_i = o)}{\mathbb{P}(W_i = W^{(0)} \mid O_i = o)} \cdot \frac{\mathbb{P}(W_i = W^{(0)})}{\mathbb{P}(W_i = W^{(j)})},$$

and estimate the conditional probabilities using logistic regression or other scalable classification methods. However, reliable estimation remains challenging in finite samples due to limited sample

sizes and sensitivity to tuning parameters (e.g., bandwidth selection), and the resulting estimation error may substantially degrade the performance.

Second, even with oracle likelihood ratios, reweighting introduces a fundamental bias-efficiency trade-off that is well known in the literature on importance weighting (e.g., Hahn (1998)). While reweighting eliminates the bias due to endogenous missingness, it can substantially increase the variance of the estimator, particularly when distributional differences across patterns are large. Appendix A.3 illustrates this trade-off in a toy example. In many cases, the bias induced by distributional differences is smaller than the additional variation introduced by reweighting, so that the unreweighted estimator can exhibit smaller mean squared error despite being biased.

Given these considerations, we recommend implementing the estimator without reweighting alongside the reweighted estimator. When the two yield similar point estimates, this provides evidence that endogenous missingness does not materially affect the conclusions, and the unreweighted estimator may be preferred due to its greater efficiency. When they differ, the reweighted estimator serves as a bias-corrected alternative, albeit with potentially larger standard errors.

6 Estimation without Fully Observed Units

This section extends our framework to settings in which no fully observed units are available. The framework developed so far relies on fully observed units to construct debiasing terms that correct for imputation error. In practice, however, there are settings where every unit has at least one missing covariate, especially when the number of covariates is large. We show that consistent estimation and valid inference remain feasible even in this challenging setting, despite the absence of complete observations.[7] For notational simplicity, we focus on the case under exogenous missingness.

When fully observed units are unavailable, there are no natural benchmarks for directly constructing the debiasing terms in equation (5). The key idea is instead to exploit the structure of the target moment and the relative completeness across partially observed units to construct debiasing sets. Formally, we focus on settings where each entry of the target moment admits an additive decomposition. Specifically, for each coordinate l, the corresponding component of the moment

[7] For simplicity, we focus on the extreme case where the number of fully observed units $N_0 = 0$. The case where $N_0 = O(1)$ is asymptotically equivalent, and the approach described below also applies when N_0 diverges.

function can be decomposed as

$$g_l(Z;\theta) = \sum_{c=1}^{C} g_{l,c}(X_{A_{l,c}}, Y;\theta),$$

where each term $g_{l,c}$ depends only on a subset of covariates $X_{A_{l,c}}$ and the outcome Y. In this setting, debiasing can be carried out component by component. For any index set $A_{l,c} \subseteq [L]$ of covariates, let

$$Q_{A_{l,c}} \coloneqq \{i : W_{ia} = 1 \text{ for all } a \in A_{l,c}\}$$

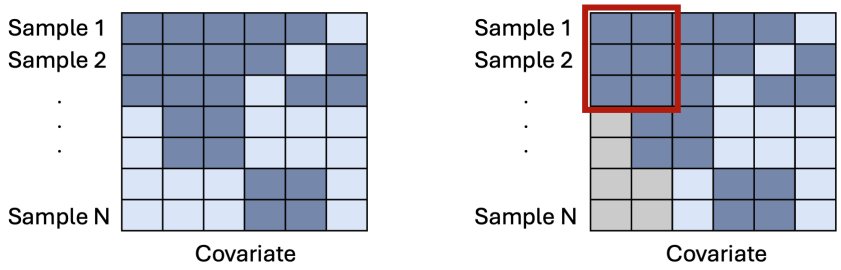
denote the set of units for which all covariates in that index set are jointly observed. For each moment component $g_{l,c}$, we use $Q_{A_{l,c}}$ as the debiasing set to correct the imputed moments of the remaining units. For identification and inference, we require that the size of each debiasing set diverges, i.e., $|Q_{A_{l,c}}| \to \infty$ for all (l,c); otherwise, the corresponding components are not consistently estimable in general without additional structural assumptions. We define $N_* \coloneqq \min_{l,c} |Q_{A_{l,c}}|$ as the minimum size across all debiasing sets, and the overall convergence rate is governed by this smallest size. This approach preserves the core insight of using observed data to correct imputation error while adapting it to settings with pervasive missingness.

We illustrate this idea in the context of linear regressions, where the target moment $g(Z;\theta) = X_i X_i^\top \theta - X_i Y_i$ can be partitioned into the components $X_{il} X_{ih} \theta_h$ for all covariate pairs (l,h) and $X_{il} Y_i$ for all l. This structure allows us to construct debiasing sets separately for each moment component. Consider, for example, how to estimate the moment $\mathbb{E}[X_{i1} X_{i2}]$ using the data with the observation pattern shown in Figure 3. In this data, the first few units are fully observed for the first two covariates, while the remaining units contain one or two missing entries for these covariates. To estimate $\mathbb{E}[X_{i1} X_{i2}]$, we first identify the units that are fully observed for this pair and use them as the debiasing set, as illustrated in the right panel of Figure 3. We then impute the missing entries for the remaining units and apply our framework with data-driven debiasing and adaptive weighting to incorporate all available data in estimating this moment. The same procedure is applied to the other components, and the resulting component-wise estimators are combined to form the final estimator.

Building on this structure, we now formalize the weighted debiased moment estimator for this setting. For linear regression, let $Q_{hl} = \{i : W_{ih} = W_{il} = 1\}$ denote the set of units for which both covariates l and h are observed, and let

$$I_{hl} = \{j \in [m] : S_j \cap Q_{hl} = S_j\}$$

Figure 3: Component-wise debiasing set construction



This figure illustrates the debiasing set for component-wise moment estimation. Each row represents a unit and each column represents a covariate, where dark blue color indicates observed entries, light blue color indicates missingness, and gray color indicates imputed values. The left panel shows the covariate observation pattern in the data. The right panel illustrates the construction of the debiasing set for estimating the moment $\mathbb{E}[X_{i1}X_{i2}]$. The red box highlights the units that are fully observed for covariates X_1 and X_2, which form the debiasing set for this moment component. Specifically, these units are used to debias the imputed covariates of the remaining units, which have missing entries in X_1, X_2, or both.

denote the set of corresponding subsets. The minimum size of the debiasing sets N_* reduces to the minimum number of jointly observed units across all covariate pairs, i.e., $N_* = \min_{h,l} |Q_{hl}|$. As in Section 3, we adopt a K-fold cross-fitting procedure where the imputation function is learned K times, each time excluding a held-out subset on which the moment function is evaluated. Under this cross-fitting procedure, $\tilde{S}_1, \ldots, \tilde{S}_K$ are disjoint equal-sized folds and $S_{j,k} = S_j \cap \tilde{S}_k$ is the set of units in S_j assigned to fold k. The resulting weighted debiased moment estimator for the l-th component is

$$
\begin{aligned}
G^{\lambda}_{N,l}(\theta) = &\sum_{h=1}^{L} \frac{1}{|Q_{hl}|} \sum_{i \in Q_{hl}} X_{il} X_{ih} \theta_h - \frac{1}{|Q_{ll}|} \sum_{i \in Q_{ll}} X_{il} Y_i \\
&+ \sum_{k=1}^{K} \sum_{h=1}^{L} \sum_{j \notin I_{hl}} \lambda_j \frac{N_j}{N} \left(\frac{1}{N_j} \sum_{i \in S_{j,k}} f^j_{-k,l}(X_i) f^j_{-k,h}(X_i) \theta_h - \frac{1}{|Q_{hl}|} \sum_{j' \in I_{hl}} \sum_{i \in S_{j',k}} f^j_{-k,l}(X_i) f^j_{-k,h}(X_i) \theta_h \right) \\
&- \sum_{k=1}^{K} \sum_{j \notin I_{ll}} \lambda_j \frac{N_j}{N} \left(\frac{1}{N_j} \sum_{i \in S_{j,k}} f^j_{-k,l}(X_i) Y_i - \frac{1}{|Q_{ll}|} \sum_{j' \in I_{ll}} \sum_{i \in S_{j',k}} f^j_{-k,l}(X_i) Y_i \right),
\end{aligned}
\tag{13}
$$

where the imputation functions $f^j_{-k}(\cdot) = f^j(\cdot; [N] \backslash \tilde{S}_k)$.

This estimator generalizes our main framework to settings without fully observed units and has several natural benchmarks as special cases. Setting $\lambda_j = 0$ for all j, estimator (13) reduces to the consistent estimator using only the observed data. When imputation perfectly recovers the missing covariates, setting $\lambda_j = 1$ for all j recovers the estimated moment with full data. More generally, appropriately chosen positive weights allow imputed data to improve estimation efficiency. While it is possible to assign different weights to each subset when they are used for estimating different

components, for simplicity we assign a common weight to each subset. The estimator and theory extend naturally to settings with component-specific weights.

The asymptotic properties of this estimator parallel those of the main framework, but with convergence rates and asymptotic variances determined by the availability of jointly observed covariates.

Theorem 4 (Asymptotic theory for estimator (13))**.** *Let $N_* = \min_{h,l} |Q_{hl}|$ denote the minimum number of jointly observed units across all covariate pairs (h,l). Suppose $N_* \to \infty$ as $N \to \infty$ and Assumption 1 holds. For any $\theta \in \Theta$, $\mathrm{Var}(X_i X_i^\top)$,$\mathrm{Var}(X_i Y_i)$, $\mathrm{Var}\big(f^j_{-k}(X_i) f^j_{-k}(X_i)^\top\big)$, and $\mathrm{Var}\big(f^j_{-k}(X_i) Y_i\big)$ are finite for any given imputation function f^j_{-k}, and $J := \mathbb{E}[X_i X_i^\top]$ is positive definite. For given weights $\lambda = (\lambda_1, \ldots, \lambda_m)$, assume $\hat{\theta}^\lambda$ uniquely solves $G_N^\lambda(\theta) = 0$ defined in (13) with probability tending to one. In addition, assume that either (1) the parameter space Θ is compact and Assumption 2.3 holds globally on Θ, or (2) $G_N^\lambda(\theta)$ is nondecreasing, i.e., $(G_N^\lambda(\theta_1) - G_N^\lambda(\theta_2))^\top(\theta_1 - \theta_2) \geq 0$ for any $\theta_1, \theta_2 \in \Theta$. Then, $\hat{\theta}^\lambda$ is consistent, i.e., $\hat{\theta}^\lambda \xrightarrow{p} \theta_0$ as $N \to \infty$. Moreover, assume the stability condition (6) holds. Then, $\hat{\theta}^\lambda$ is asymptotically normal:*

$$\sqrt{N_*}(\hat{\theta}^\lambda - \theta_0) \to \mathcal{N}\left(0, J^{-1} \sum_{j=1}^{m} \mathrm{Cov}(\varphi_{j,i}) J^{-1}\right).$$

where the l-th component of $\varphi_{j,l}$ is defined as

$$\varphi_{j,i,l} = \sum_{h=1}^{L} I(j \in I_{hl}) r_{*,j,hl} \left(X_{il} X_{ih} \theta_{0,h} - \sum_{j' \notin I_{hl}} \lambda_{j'} \frac{N_{j'}}{N} \bar{f}_l^{j'}(X_i) \bar{f}_h^{j'}(X_i) \theta_{0,h} \right) + I(j \notin I_{hl}) \lambda_j r_{*,j} \bar{f}_l^j(X_i) \cdot$$
$$\bar{f}_h^j(X_i) \theta_{0,h} - \left[I(j \in I_{ll}) r_{*,j,ll} \left(X_{il} Y_i - \sum_{j' \notin I_{ll}} \lambda_{j'} \frac{N_{j'}}{N} \bar{f}_l^{j'}(X_i) Y_i \right) + I(j \notin I_{ll}) \lambda_j r_{*,j} \bar{f}_l^j(X_i) Y_i \right].$$

and $r_{*,j,hl} = \lim_{N\to\infty} \sqrt{N_* N_j}/|Q_{hl}|$ and $r_{*,j} = \lim_{N\to\infty} \sqrt{N_* N_j}/N$.

Theorem 4 establishes consistency and asymptotic normality of the estimator $\hat{\theta}^\lambda$. To ensure consistent estimation, we require that each pair of covariates is jointly observed by sufficiently many units, that is, $N_* \to \infty$ as $N \to \infty$. The estimator converges at rate $\sqrt{N_*}$, determined by the smallest debiasing set. The asymptotic variance has a structure analogous to that in Theorem 2, with contributions from both observed and imputed data, demonstrating the efficiency trade-off from including the imputed values.

This extension substantially broadens the applicability of our framework. Even in settings with

pervasive missingness and no fully observed units, our approach provides consistent and efficient estimation with valid inference by exploiting the relative information content across observation patterns. While the exposition focuses on linear regression for clarity, this approach applies to other specifications whose moment functions admit a finite additive decomposition into components that depend on low-dimensional subsets of covariates, including polynomial regressions and models with interaction terms.

Furthermore, this approach also applies when fully observed units are available. Relative to the baseline approach in Section 3, where debiasing relies on the N_0 fully observed units and yields a $\sqrt{N_0}$ convergence rate, the component-wise approach allows debiasing sets to be formed separately for each moment component. As a result, when each component depends only on a low-dimensional subset of covariates, the corresponding debiasing sets can be substantially larger, leading to a faster convergence rate governed by N_*.

7 Simulation

This section demonstrates the superior performance of the proposed framework through a comprehensive set of simulation studies. We compare our approach to two commonly used but naive approaches for handling missing covariates: (1) complete-case analysis, which uses only the fully observed units, and (2) naive imputation, which treats imputed values as if they were observed.

We generate a covariate panel $\mathbf{X} = [X_1, \ldots, X_N]^\top \in \mathbb{R}^{N \times L}$ using a latent factor model with k latent factors, and generate outcomes Y_i through a linear model:

$$X_{il} = \Lambda_i^\top F_l + e_{il}, \quad Y_i = X_i^\top \theta_0 + \epsilon_i,$$

where the factors $F_l \overset{i.i.d.}{\sim} \mathcal{N}(0, I_k)$, loadings $\Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_k)$, idiosyncratic covariate noise $e_{il} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_e^2)$, and outcome noise $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. This latent factor structure captures the correlation patterns commonly found in high-dimensional datasets. The covariate vectors X_i exhibit heterogeneous observation patterns. Specifically, the sample contains 20 distinct observation patterns, consisting of one fully observed pattern and $m = 19$ partially observed patterns with different proportions and locations of missing covariates. Each partially observed unit is randomly assigned to one of these patterns. In our design, the first half of the partially observed patterns contains 60% observed covariates, while the second half contains proportions $p \in \{40\%, 60\%, 80\%\}$ of observed covariates. Table 1 provides a detailed description of the data-generating process and observation

Table 1: Comparison of estimators for different observation proportions p

		$k_0 = 5, \hat{k} = 5$ correct imputation model				$k_0 = 5, \hat{k} = 3$ misspecified imputation model			
p	Method	$\bar{\lambda}$	|Bias|	Var	MSE	$\bar{\lambda}$	|Bias|	Var	MSE
40%	**Our method**	0.78	**0.008**	**0.198**	**0.200**	0.52	**0.007**	**0.221**	**0.223**
	Complete subset	0.0	**0.007**	0.293	0.295	0.0	**0.007**	0.293	0.295
	Naive imputation	-	0.041	0.380	0.416	-	0.165	3.144	3.733
60%	**Our method**	0.90	**0.008**	0.170	**0.172**	0.58	**0.007**	**0.201**	**0.203**
	Complete subset	0.0	**0.007**	0.293	0.295	0.0	**0.007**	0.293	0.295
	Naive imputation	-	0.020	**0.160**	**0.171**	-	0.153	2.259	2.786
80%	**Our method**	0.92	**0.006**	0.110	0.111	0.69	**0.006**	**0.136**	**0.137**
	Complete subset	0.0	**0.007**	0.293	0.295	0.0	**0.007**	0.293	0.295
	Naive imputation	-	0.011	**0.090**	**0.094**	-	0.117	1.158	1.485

This table reports the average absolute bias (|Bias|), total variance (Var), and mean squared error (MSE) for three methods: our method, complete-case analysis (complete subset), and naive use of imputation (naive imputation). For our method, we use $K = 10$ folds for cross-fitting and $B = 50$ bootstrap samples for the weight selection. Bold numbers indicate the (near) best performance across models. The data-generating process and covariate observation patterns are as follows: $\sigma_e = 1$, $\sigma_\epsilon = 3$, and the true regression coefficient is $\theta_0 = \mathbf{1}_L$ (vector of ones). We assume $m = 19$ partially observed patterns where the first half contains 60% observed covariates and the second half contains p observed covariates with $p \in \{40\%, 60\%, 80\%\}$. Each partially observed unit is randomly assigned to one of the m partially observed patterns. For each pattern, the set of observed covariate coordinates is drawn uniformly at random, subject to the observation proportion, and the remaining coordinates are treated as missing. We set $N = 5000$, $N_0 = 500$, $L = 20$, and run 200 simulations for each setup.

patterns. The target parameter is the regression coefficient vector $\theta_0 \in \mathbb{R}^L$, defined by the moment condition $G(\theta) = \mathbb{E}[X_i X_i^\top \theta - X_i Y_i] = 0$.

Regarding imputation methods, we employ the factor model imputation with $\hat{k}$ latent factors for both our estimator and the naive imputation approach. To evaluate the effect of imputation quality, we consider both correctly specified imputation models ($\hat{k} = k$) and misspecified models ($\hat{k} < k$). Under correct specification, the imputed values are asymptotically unbiased, whereas misspecification introduces bias and additional variance into the imputed covariates.

Table 1 compares the performance of the three estimators across different settings, reporting average absolute bias (|Bias|), total variance (Var), and mean squared error (MSE) over $n = 200$ simulations. These metrics are respectively defined as

$$|\text{Bias}| = \frac{1}{L}\sum_{l=1}^{L}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{il} - \theta_{0l}\right|, \quad \text{Var} = \frac{1}{n}\sum_{i=1}^{n}\|\hat{\theta}_i - \bar{\theta}\|^2, \quad \text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\|\hat{\theta}_i - \theta_0\|^2,$$

where $\bar{\theta} = n^{-1} \sum_{i=1}^{n} \hat{\theta}_i$. For our method, we impose the constraint $\lambda_j \in [0, 1]$ and report the average optimized weight $\bar{\lambda} \coloneqq m^{-1} \sum_{j=1}^{m} \lambda_j^*$.[8]

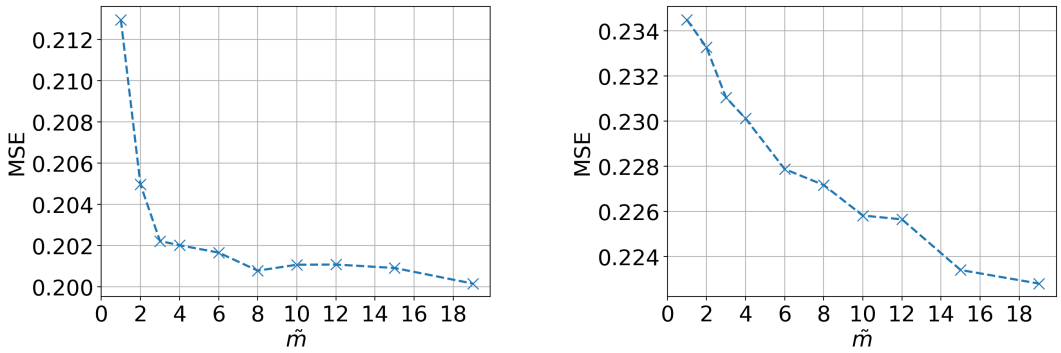
The results in Table 1 demonstrate that our estimator performs well across all settings, consistently achieving low bias and low variance, whereas the benchmark methods are either biased, less efficient, or both. Compared to the complete-case analysis, our method delivers comparable unbiasedness but substantially higher efficiency by appropriately incorporating information from partially observed units. The efficiency gains are particularly pronounced when imputation quality is high or the proportion of observed covariates is large. For example, when $p = 80\%$ and the imputation model is correctly specified, our method achieves an MSE of 0.111, compared to 0.295 for the complete-case analysis. Importantly, even under misspecified imputation and substantial missingness, our method continues to outperform the complete-case analysis through its adaptive weighting scheme. The magnitude of these efficiency gains depends on both imputation quality and the extent of missingness.

Relative to the naive imputation method, our approach exhibits strong robustness to imputation errors. While naive imputation can perform competitively when imputation accuracy is high and missingness is limited, its performance deteriorates significantly under poor imputation quality. For example, when $p = 40\%$ and the imputation model is misspecified, the MSE of naive imputation increases to 3.733. In contrast, our estimator automatically shrinks the optimized weights toward zero, effectively downweighting the unreliable imputed values. Notably, our approach can outperform naive imputation even when the imputation model is correctly specified. For example, with $p = 40\%$ under correct imputation, our approach achieves a lower variance than naive imputation despite explicitly accounting for imputation uncertainty. This improvement arises because the optimal weighting scheme effectively differentiates imputed observations by reliability, rather than treating them uniformly.

We next evaluate the performance of the data-driven grouping procedure introduced in Sections 3.2 and A.1. In the simulations above, each partially observed pattern is assigned its own weight, which may become computationally costly and statistically unstable when the number of partially observed patterns m is large. The grouping approach mitigates this issue by partitioning the m patterns into $\tilde{m}$ groups and assigning a common weight to each group. This approach preserves the unbiasedness of the estimator and affects only its efficiency through dimension reduction in the weighting scheme. Figure 4 shows the MSE of the grouped-weight estimator as a function of the

[8] Allowing unconstrained weights can yield slightly lower variance and mean squared error in some settings. However, to reflect high-stakes applications where imposing weight constraints is desirable, we report the more conservative results based on constrained weights.

Figure 4: Performance of the data-driven grouping method



(a) Correct imputation model **(b)** Misspecified imputation model

These figures report the mean squared error (MSE) of the grouped-weight debiased estimator as a function of the number of groups $\tilde{m}$. We consider the simulation design in Table 1 with $m = 19$ distinct partially observed patterns and $p = 40\%$ observed covariates in the second half of the partially observed data. Panel (a) uses the correctly specified imputation model, and panel (b) uses the misspecified imputation model. For a given $\tilde{m}$, we partition the m patterns into $\tilde{m}$ groups following the procedure in Section A.1 and construct the grouped-weight debiased estimator as in Section 3.2. The case $\tilde{m} = 0$, which corresponds to the estimator without weights for imputed data, yields an MSE of 0.218 in panel (a) and 0.305 in panel (b) and is omitted from the plots to avoid distorting the scale. We run 200 simulations for each setup.

number of groups $\tilde{m}$ for the setting with $p = 40\%$, under both correctly specified and misspecified imputation models. When $\tilde{m} = 19$, the estimator coincides with the original one without grouping, while smaller values of $\tilde{m}$ correspond to coarser groupings.

The results suggest that a moderate number of groups ($\tilde{m} \leq 10$) captures most of the efficiency gains from pattern-specific weighting, under both correct and misspecified imputation models. The MSE decreases rapidly as $\tilde{m}$ increases from very small values and then flattens, indicating diminishing efficiency gains from further refinement. Under correct imputation, the efficiency gains diminish particularly quickly. In this setting, the imputation errors cluster naturally according to the proportion of observed covariates; therefore, even a small number of groups captures the relevant heterogeneity in imputation quality.

Finally, we study the benefits of using heterogeneous weights in our estimator. Table 2 considers a setting with extreme missingness heterogeneity. Specifically, the data contains two partially observed patterns: one with 5% missing covariates and the other with 70% missingness. In addition to the benchmark methods, we include a single-weight variant of our estimator that assigns a common optimized weight to all partially observed units.

Table 2 shows that even the single-weight variant improves on the naive approaches. However, when observation patterns exhibit substantial heterogeneity, allowing different weights provides large

Table 2: Comparison of estimators under extreme missingness heterogeneity

	$k_0 = 5, \hat{k} = 5$ correct imputation model				$k_0 = 5, \hat{k} = 3$ misspecified imputation model			
	$\bar{\lambda}$	|Bias|	Var	MSE	$\bar{\lambda}$	|Bias|	Var	MSE
Our method	0.55	**0.005**	**0.152**	**0.153**	0.53	**0.006**	**0.162**	**0.163**
Single-weight variant	0.57	**0.006**	0.229	0.231	0.36	**0.007**	0.244	0.246
Complete subset	0.0	**0.007**	0.293	0.295	0.0	**0.007**	0.293	0.295
Naive imputation	-	0.059	0.735	0.834	-	0.078	2.617	2.831

This table reports the average absolute bias (|Bias|), total variance (Var), and mean squared error (MSE) for four methods: our method (with different optimized weights on the two patterns), a variant of our method that assigns a single optimized weight to all partially observed units (Single-weight variant), complete-case analysis (complete subset), and naive use of imputation (naive imputation). The data-generating process follows Table 1 with modified observation patterns: two partially observed patterns with 5% and 70% missing covariates, respectively. We set $N = 5000$, $N_0 = 500$, $L = 20$, and run 200 simulations for each setup.

additional efficiency gains. This improvement arises because the single-weight estimator necessarily compromises across observation patterns, whereas heterogeneous weights allow the estimator to fully exploit high-quality imputations while protecting against low-quality ones.

Overall, these simulation results demonstrate three key findings. First, accounting for imputation uncertainty through explicit debiasing is essential for reliable inference. Second, adaptive weighting provides robustness to imputation misspecification while improving efficiency. Third, the data-driven grouping strategy offers a practical and effective approximation when the number of observation patterns is large. Additional simulation results in Appendix A.2 confirm that these conclusions are robust to alternative designs and noise levels.

8 Empirical Study: Carbon Emissions and Stock Returns

In this section, we apply our framework to a fundamental question in environmental finance: Are firms' carbon emissions correlated with their stock returns after controlling for firm fundamentals? This question has attracted considerable attention since Bolton and Kacperczyk (2021) and has yielded contradictory findings due to the extensive missingness in the emissions data. Bolton and Kacperczyk (2021) find a positive correlation between stock returns and carbon emissions when using vendor-imputed data, but subsequent work by Aswani, Raghunandan, and Rajgopal (2024) finds no correlation when using only the subset of fully observed data. These contradictory findings suggest that empirical conclusions can hinge on how missing data are handled. Our framework provides

an approach to resolve this debate by properly incorporating the partially observed samples with imputed data.

8.1 Data

Our dataset contains 205,267 firm-month observations from 2,890 unique firms over the period 2006-2023. We construct this dataset by combining carbon emissions data from the Trucost Environmental dataset with stock returns and firm fundamentals from the CRSP–Compustat merged dataset, both accessed through WRDS. To ensure accurate firm identification across data sources, we merge the datasets using both ticker and gvkey identifiers.[9]

The emissions data contain extensive missingness across all three emission scopes. Specifically, 62% of observations are missing for Scope 1 emissions (direct emissions from owned facilities), 64% for Scope 2 emissions (emissions from purchased energy), and 91% for Scope 3 emissions (emissions generated along the value chain). We identify an emissions observation as missing when Trucost labels it as "estimated" rather than "reported". As a result, only self-reported emissions are treated as observed data in our analysis.

In addition to emissions, we include a set of firm characteristics that are standard in the empirical asset pricing literature. These variables include firm size, leverage, the book-to-market ratio, return on equity (ROE), growth rates (EPSGR and SalesGR), capital intensity measured by log property, plant, and equipment (LogPPE) and investment-to-assets (I/A), industry concentration (HHI), and risk measures including volatility, momentum, and beta. Most variables are directly derived from CRSP–Compustat. Beta is calculated using S&P 500 returns as the market proxy, and HHI is constructed from Compustat Business Segment data by summing squared sales shares within each industry-year. Tables A.4 and A.5 report detailed variable definitions and summary statistics. Following the literature, we assign annual variables to all months within the corresponding fiscal year to construct the firm-month observations.

8.2 Regression Analysis

Following Bolton and Kacperczyk (2021) and Aswani, Raghunandan, and Rajgopal (2024), we examine the relationship between carbon emissions and stock returns using the cross-sectional

[9]We use gvkey to prevent erroneous matches when tickers are reassigned to different firms. We have identified approximately 300 instances where a single ticker corresponds to multiple CUSIPs across our sample period (though never overlapping within a month), typically occurring around major corporate events such as IPOs or mergers. In these cases, we split the ticker history at the CUSIP change point and treat them as separate firms. CUSIP matching is infeasible because Trucost data in WRDS does not provide CUSIP identifiers.

regression model:

$$RET_{it} = \alpha + \beta_1 \cdot LOG(Emission)_{it} + \beta_2^\top \cdot Controls_{it} + \delta_{g(i)} + \gamma_t + \epsilon_{it}, \tag{14}$$

where RET_{it} denotes the monthly stock return of firm i at time t, and *Emission* refers to either Scope 1, Scope 2, or Scope 3 carbon emissions. The control vector includes the firm characteristics described in Section 8.1. The time fixed effects γ_t and the industry fixed effects $\delta_{g(i)}$, where $g(i)$ denotes the industry of firm i, control for common time trends and industry-level heterogeneity. Our target parameter is the emissions coefficient β_1, which captures the carbon-return relationship after controlling for firm fundamentals and fixed effects.

The extensive missingness in the emissions data poses a fundamental challenge for estimating this regression. Previous studies address this challenge using different approaches. Bolton and Kacperczyk (2021) employ a naive imputation approach that treats the vendor-estimated emissions as observed data. Aswani, Raghunandan, and Rajgopal (2024) consider the complete-case analysis that conducts the analysis on the subset of firm-months with reported emissions. In contrast to these approaches, our framework incorporates all available data while explicitly accounting for imputation uncertainty through debiasing and adaptive weighting.

Before implementing our approach, we first replicate the main findings from the existing literature using the two naive approaches: naive imputation and complete-case analysis. To do so, we remove the time and industry fixed effects through a two-way within transformation of both the outcome and covariates, as detailed in Appendix A.5, and then conduct standard regression analysis. We account for time-series dependence when estimating standard errors. The first two rows of Table 4 present the corresponding results. Consistent with prior studies, naive imputation yields significantly positive coefficients for all three emission scopes, whereas complete-case analysis finds no significant relationship. Minor differences relative to the original papers arise from our extended sample period through 2023.

We then implement our approach. As in the naive approaches, we begin by removing the fixed effects from the data. Following Sections 3 and A.2, we partition the data according to emissions data availability. Specifically, the fully observed subset S_0 contains firm-months with reported emissions. The remaining observations with missing emissions are divided into two subsets based on the availability of temporal information:

- S_1: firm-months for which the firm reports emissions in other time periods;

- S_2: firm-months from firms that never report the corresponding emission scope in any period.

We construct separate imputation models for S_1 and S_2. For S_1, the imputation model incorporates within-firm temporal emissions in addition to the cross-sectional firm characteristics.[10] For S_2, where no temporal emissions data are available, imputation relies only on the cross-sectional characteristics. In constructing the imputation models, we leverage additional firm fundamentals, including log sales and total assets, beyond the control variables included in the regression analysis. We select the imputation method and associated hyperparameters for each subset-scope combination using five-fold cross-validation at the firm level. Specifically, we consider ridge regression, random forest, and factor model imputation, and select the method that maximizes out-of-sample R^2. Additional implementation details are provided in Appendix A.5.

Table 3 reports the cross-validation results for different imputation methods. For S_1, ridge regression performs best across all three emission scopes, achieving out-of-sample R^2 values exceeding 0.8. For S_2, where temporal information is unavailable, random forest performs best with out-of-sample R^2 around 0.4–0.5. These results indicate that while cross-sectional firm characteristics are informative, imputation accuracy improves substantially when temporal information is available. This heterogeneity in imputation quality motivates differential weighting of imputed observations in the downstream estimation.

After selecting the imputation methods, we train the models within our cross-fitting framework and construct the weighted debiased moment estimator. Since the debiasing sets differ across S_1 and S_2, we include a weight-normalization term as introduced in Section A.2 and obtain the estimator

$$G_N^{\lambda}(\theta) = \frac{1}{|S_0|}\sum_{(i,t)\in S_0} g_\theta(X_{it}, Y_{it}) - \sum_{j=1}^{2}\lambda_j \frac{|S_j|}{N}\left[\frac{1}{|S_0|}\sum_{(i,t)\in S_0} g_\theta(X_{it}, Y_{it}) - \frac{1}{|S_{0,j}|}\sum_{(i,t)\in S_{0,j}} g_\theta(X_{it}, Y_{it})\right]$$
$$+ \sum_{k=1}^{K}\sum_{j=1}^{2}\lambda_j \frac{|S_j|}{N}\left(\frac{1}{|S_j|}\sum_{(i,t)\in S_{j,k}} g_\theta(f_{-k}^{j}(X_i), Y_{it}) - \frac{1}{|S_{0,j}|}\sum_{(i,t)\in S_{0,j,k}} g_\theta(f_{-k}^{j}(X_i), Y_{it})\right),$$

where N denotes the number of all firm-month observations, and for the linear regression specification, the moment function is $g_\theta(x, y) = xx^\top\theta - xy$. We determine the subset weights $\lambda = (\lambda_1, \lambda_2)$ by minimizing the asymptotic variance of the emissions coefficient $\hat{\beta}_1$ and impose the constraints $\lambda_j \in [0, 1]$ to avoid over-reliance on the imputed data. The resulting optimized weights are $\hat{\lambda} = (1.00, 0.24)$

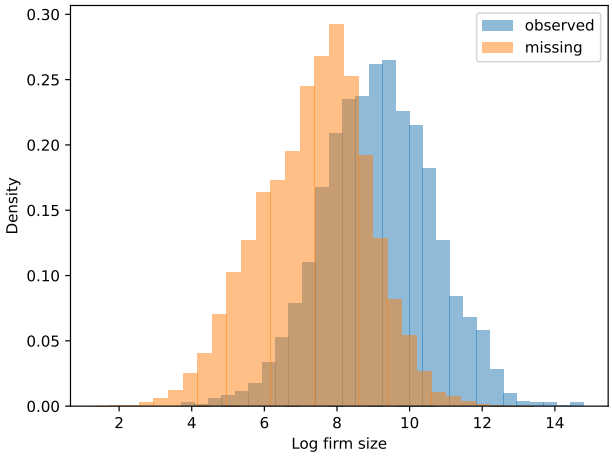
[10] For within-firm temporal emissions, we use the observation that is temporally closest to the target period, selecting past values when past and future observations are equidistant. We note, however, that temporal proximity does not necessarily imply value similarity, as emissions can exhibit substantial volatility around a baseline level.

Table 3: Imputation performance of different methods

Subset	Method	Out-of-sample R^2 Scope 1	Scope 2	Scope 3
S_1: Temporal emissions available	Ridge	**0.832**	**0.839**	**0.862**
	RF	0.828	0.836	0.838
	FM	0.799	0.792	0.841
S_2: Neither past nor future	Ridge	0.370	0.397	0.337
	RF	**0.419**	**0.476**	**0.440**
	FM	0.212	0.212	0.081

This table reports the out-of-sample R^2 of different imputation methods for each subset-scope combination. Methods include ridge regression (Ridge), random forest (RF), and factor model (FM). All methods are evaluated using five-fold cross-validation with optimally tuned hyperparameters. We use bold numbers to indicate best performance. For subset S_1, the numbers of observations are 49,084, 51,053, and 18,494 for Scope 1, Scope 2, and Scope 3 emissions, respectively. For subset S_2, the corresponding numbers are 77,256, 80,921, and 169,116.

Figure 5: Firm size distributions by emissions missingness



This figure shows the density distributions of log firm size for observations with and without Scope 1 emissions data.

for Scope 1, $\hat{\lambda} = (1.00, 0.22)$ for Scope 2, and $\hat{\lambda} = (1.00, 0.41)$ for Scope 3 emissions. These weights reflect imputation quality, as in all cases, the imputed data in S_1 receive substantially higher weights than those in S_2.

We additionally implement the reweighted estimator from Section 5 to account for potential endogenous missingness of emissions data. As noted in the literature, larger firms are more likely to report their emissions. Figure 5 illustrates this difference between reporting and non-reporting firms in our sample. We apply the reweighted estimator to assess whether accounting for these

Table 4: Comparison of methods: Carbon emissions and stock returns

	Scope 1	Scope 2	Scope 3
Naive imputation			
Coefficient $\hat{\beta}_1$	0.128***	0.131***	0.328***
SE	(0.028)	(0.034)	(0.047)
Complete subset			
Coefficient $\hat{\beta}_1$	0.040	-0.006	0.156
SE	(0.032)	(0.035)	(0.118)
Our method			
Coefficient $\hat{\beta}_1$	0.023	-0.041	-0.175**
SE	(0.024)	(0.027)	(0.086)
Reweighted complete subset			
Coefficient $\hat{\beta}_1$	-0.011	-0.091	-0.418*
SE	(0.055)	(0.058)	(0.243)
Reweighted our method			
Coefficient $\hat{\beta}_1$	0.025	-0.037	-0.361
SE	(0.046)	(0.049)	(0.230)
Observation Number N			
Imputed data	205,111	205,236	205,267
Complete subset	78,927	73,293	17,657

*Significance levels: *** p<0.01, ** p<0.05, * p<0.1.*

This table compares regression coefficients $\hat{\beta}_1$ for log carbon emissions across different estimation methods. Naive imputation fills missing emissions with vendor-estimated values and treats them as observed data. The complete-case analysis (complete subset) uses only firm-month observations with fully observed emissions. Our method uses all observations while accounting for imputation uncertainty through bias correction and optimal weighting. The reweighted versions apply likelihood ratio reweighting to account for endogenous missingness based on firm size. Standard errors (SE) are clustered at the firm level and computed using the estimator obtained under each method. We use $B = 50$ in the bootstrap approach for variance estimation in our approaches.

size differences affects our empirical results. Specifically, we construct likelihood ratios based on within-transformed firm size using Gaussian kernel density estimation and re-optimize the subset weights to minimize the asymptotic variance of the emissions coefficient.[11]

Table 4 reports the results of our approaches. For comparison, we also report a reweighted complete-case analysis, which is a special case of our reweighted approach with zero weight placed on the imputed data. There are two main takeaways from Table 4. First, methodological choices can fundamentally change the economic conclusions. Whereas the naive imputation method finds

[11] When estimating the optimal weights in our method, we use the complete-case estimator to compute the asymptotic variance matrix Σ^{λ}. For the reweighted version of our method, we instead use the reweighted complete-case estimator in computing the corresponding asymptotic variance. In both cases, we conduct a grid search over $[0, 1]^2$ to determine the optimized weights. For the reweighted version, the resulting optimized weights are $\hat{\lambda} = (1.00, 0.75)$ for Scope 1, $\hat{\lambda} = (1.00, 0.22)$ for Scope 2, and $\hat{\lambda} = (1.00, 0.19)$ for Scope 3.

significant positive coefficients for all emission scopes, both variants of our approach suggest no significant relationship for Scopes 1 and 2 and a negative coefficient for Scope 3. A similar finding appears in the reweighted complete-case analysis, which also yields a negative coefficient for Scope 3. Taken together, these results suggest that the previously documented "carbon premium" may be a methodological artifact rather than a market phenomenon. Furthermore, the difference between the unreweighted and the reweighted approaches for Scope 3 emissions indicates that the larger firms, which are more likely to report emissions, may have different emission-return relationships than the broader population.

Second, correcting for endogenous missingness through reweighting substantially increases the standard errors, especially for Scope 3 emissions. As discussed in Section 5, reweighting introduces additional uncertainty through both the reweighting itself and the estimation of the density ratios. This reflects a fundamental bias-variance trade-off: our original approach yields smaller standard errors but may be biased when missingness is endogenous, while the reweighted approach mitigates this bias at the cost of larger variance. In Table 4, the reweighted approaches continue to yield negative coefficients for Scope 3 emissions, suggesting that the bias related to firm size is unlikely to be substantial in this setting. In this case, the original estimator may therefore be preferable, as its smaller standard errors enable detection of relationships that would be obscured by the larger standard errors of the reweighted approach.

Overall, these results indicate that treating imputed values as observed data can lead to spurious findings and misleading conclusions. Our framework, which explicitly accounts for imputation uncertainty through bias correction and adaptive weighting, finds that the emissions from firm operations (Scopes 1 and 2) are not correlated with stock returns, but that the value chain emissions (Scope 3) are negatively correlated with returns.

9 Conclusion

This paper develops a general framework for conducting valid statistical inference with missing covariate data. Our approach addresses the fundamental disconnect between sophisticated imputation methods and their reliable use in downstream statistical analyses. By combining bias correction based on observed data with adaptive weighting that reflects imputation reliability, we enable researchers to leverage all available data while maintaining statistical efficiency and validity.

This framework provides an automatic pipeline for empirical researchers to estimate models

from datasets with missing covariates. The process is simple: researchers input their incomplete dataset with complex observation patterns and choose an imputation method, ranging from factor-model imputations to machine learning approaches. Our framework then corrects imputation bias, adaptively weights observations based on imputation quality across different observation patterns, and delivers estimates together with inferential theory. While deriving the underlying theory is challenging, this pipeline is remarkably easy to implement and allows researchers to leverage all available data with rigorous statistical guarantees.

In simulations, we confirm that our approach consistently outperforms both complete-case analysis and naive imputation across various settings. Our empirical study in environmental finance further demonstrates the practical importance of the framework. By properly accounting for the uncertainty induced by imputation, our approach suggests that the previously reported carbon premium may be a statistical artifact rather than an economic reality.

References

Abrevaya, J., and S. G. Donald (2017): "A GMM approach for dealing with missing data on regressors," *Review of Economics and Statistics*, 99(4), 657–662.

Angelopoulos, A. N., S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic (2023): "Prediction-powered inference," *arXiv preprint arXiv:2301.09633*.

Angelopoulos, A. N., J. C. Duchi, and T. Zrnic (2023): "Ppi++: Efficient prediction-powered inference," *arXiv preprint arXiv:2311.01453*.

Aswani, J., A. Raghunandan, and S. Rajgopal (2024): "Are carbon emissions associated with stock returns?," *Review of Finance*, 28(1), 75–106.

Bai, J., and S. Ng (2021): "Matrix completion, counterfactuals, and factor analysis of missing data," *Journal of the American Statistical Association*, 116(536), 1746–1763.

Bayle, P., A. Bayle, L. Janson, and L. Mackey (2020): "Cross-validation confidence intervals for test error," *Advances in Neural Information Processing Systems*, 33, 16339–16350.

Berrett, T. B. (2024): "Efficient estimation with incomplete data via generalised ANOVA decomposition," *arXiv preprint arXiv:2409.05729*.

Bolton, P., and M. Kacperczyk (2021): "Do investors care about carbon risk?," *Journal of financial economics*, 142(2), 517–549.

Bousquet, O., and A. Elisseeff (2002): "Stability and generalization," *Journal of Machine Learning Research*, 2(Mar), 499–526.

Bryzgalova, S., S. Lerner, M. Lettau, and M. Pelger (2025): "Missing financial data," *The Review of Financial Studies*, 38(3), 803–882.

Bühlmann, P., and S. van de Geer (2011): *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Berlin.

Cahan, E., J. Bai, and S. Ng (2023): "Factor-based imputation of missing values and covariances in panel data of large dimensions," *Journal of Econometrics*, 233(1), 113–131.

Candès, E., and B. Recht (2012): "Exact matrix completion via convex optimization," *Communications of the ACM*, 55(6), 111–119.

Chakrabortty, A., and T. Cai (2018): "Efficient and adaptive linear regression in semi-supervised settings," *The Annals of Statistics*, 46(4), 1541–1572.

Chen, X., T. McCormick, B. Mukherjee, and Z. Wu (2025): "A Unified Framework for Inference with General Missingness Patterns and Machine Learning Imputation," *arXiv preprint arXiv:2508.15162*.

Chen, Y., J. Fan, C. Ma, and Y. Yan (2019): "Inference and uncertainty quantification for noisy matrix completion," *Proceedings of the National Academy of Sciences*, 116(46), 22931–22937.

Chernozhukov, V., C. Hansen, Y. Liao, and Y. Zhu (2023): "Inference for low-rank models," *The Annals of Statistics*, 51(3), 1309–1330.

Choi, J., and M. Yuan (2024): "Matrix completion when missing is not at random and its applications in causal panel data models," *Journal of the American Statistical Association*, (just-accepted), 1–24.

Deng, S., Y. Ning, J. Zhao, and H. Zhang (2024): "Optimal and safe estimation for high-dimensional semi-supervised learning," *Journal of the American Statistical Association*, 119(548), 2748–2759.

Donders, A. R. T., G. J. van der Heijden, T. Stijnen, and K. G. Moons (2006): "Review: A gentle introduction to imputation of missing values," *Journal of Clinical Epidemiology*, 59(10), 1087–1091.

Duan, J., M. Pelger, and R. Xiong (2024a): "Factor Analysis for Causal Inference on Large Non-Stationary Panels with Endogenous Treatment," *Available at SSRN 4823360*.

——— (2024b): "Target PCA: Transfer learning large dimensional panel data," *Journal of Econometrics*, 244(2), 105521.

Freyberger, J., B. Höppner, A. Neuhierl, and M. Weber (2024): "Missing data in asset pricing panels," *The Review of Financial Studies*, p. hhae003.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002): *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, pp. 315–331.

Hansen, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the econometric society*, pp. 1029–1054.

Ji, W., L. Lei, and T. Zrnic (2025): "Predictions as surrogates: Revisiting surrogate outcomes in the age of ai," *arXiv preprint arXiv:2501.09731*.

Jin, S., K. Miao, and L. Su (2021): "On factor models with random missing: Em estimation, inference, and cross validation," *Journal of Econometrics*, 222(1), 745–777.

Jones, M. P. (1996): "Indicator and stratification methods for missing explanatory variables in multiple linear regression," *Journal of the American statistical association*, 91(433), 222–230.

Kluger, D. M., K. Lu, T. Zrnic, S. Wang, and S. Bates (2025): "Prediction-Powered Inference with Imputed Covariates and Nonuniform Sampling," *arXiv preprint arXiv:2501.18577*.

Little, R. J. (1986): "Survey nonresponse adjustments for estimates of means," *International Statistical Review/Revue Internationale de Statistique*, pp. 139–157.

Mazumder, R., T. Hastie, and R. Tibshirani (2010): "Spectral regularization algorithms for learning large incomplete matrices," *The Journal of Machine Learning Research*, 11, 2287–2322.

Miao, J., X. Miao, Y. Wu, J. Zhao, and Q. Lu (2023): "Assumption-lean and data-adaptive post-prediction inference," *arXiv preprint arXiv:2311.14220*.

Negahban, S., and M. J. Wainwright (2011): "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, 39(2), 1069–1097.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011): "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, 12, 2825–2830.

Van der Vaart, A. W. (2000): *Asymptotic statistics*, vol. 3. Cambridge university press.

Van Ness, M., T. M. Bosschieter, R. Halpin-Gregorio, and M. Udell (2023): "The missing indicator method: From low to high dimensions," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5004–5015.

Wager, S., and S. Athey (2018): "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 113(523), 1228–1242.

Wang, S., T. H. McCormick, and J. T. Leek (2020): "Methods for correcting inference based on outcomes predicted by machine learning," *Proceedings of the National Academy of Sciences*, 117(48), 30266–30275.

Wasserman, L., and J. Lafferty (2007): "Statistical analysis of semi-supervised regression," *Advances in Neural Information Processing Systems*, 20.

Xiong, R., and M. Pelger (2023): "Large dimensional latent factor modeling with missing observations and applications to causal inference," *Journal of Econometrics*, 233(1), 271–301.

Xu, Q., L. Testa, J. Lei, and K. Roeder (2025): "Blockwise Missingness meets AI: A Tractable Solution for Semiparametric Inference," *arXiv preprint arXiv:2509.24158*.

Zhao, S., and E. Candès (2025): "Imputation-Powered Inference," *arXiv preprint arXiv:2509.13778*.

Appendix

A.1 Grouping of Observation Patterns

In this section, we provide the details of the grouping approach introduced in Section 3.2. The goal of this approach is to construct a practical approximation to the optimal weighting scheme characterized in Corollary 1 when the number of distinct observation patterns is large. To achieve this, we group observation patterns that have similar optimal weights and assign them a common group-level weight.

Corollary 1 shows that the optimal weights are proportional to the error in the moment functions induced by imputation. Under moment smoothness conditions, this error can be controlled by the imputation error in the covariates. Based on this relationship, we propose to use the mean squared imputation error as a proxy for the imputation error and to construct the grouping based on this measure.

More specifically, we proceed as follows. We first specify a target number of groups $\tilde{m}$. We then estimate the mean squared imputation error for each observation pattern j. To do so, we partition the data into training and validation sets, estimate the imputation function f^j on the training set, apply it to the fully observed units in the validation set, and compute the imputation error

$$\mathcal{E}_j = \frac{1}{|S_0^{val}|} \sum_{i \in S_0^{val}} \left\| f^j_{train}(X_i, W^{(j)}) - X_i \right\|^2,$$

where S_0^{val} denotes the validation subset of the fully observed units. Once the imputation errors $\{\mathcal{E}_1, \ldots, \mathcal{E}_m\}$ are computed, we group the observation patterns with similar errors using a clustering algorithm, such as k-means or hierarchical clustering. In the simulations in Section 7, we implement this step using the k-means clustering algorithm.

For example, consider a data set with 10,000 units and 20 covariates, where 10 covariates contain missing entries and the sample exhibits 100 distinct partially observed patterns. Suppose the imputation errors satisfy $\Delta_{il} \overset{i.i.d.}{\sim} \mathcal{N}(\mu_\Delta, \sigma^2_\Delta)$. For an observation pattern j with d_j missing entries, the corresponding imputation error satisfies $\mathcal{E}_j \approx d_j(\mu^2_\Delta + \sigma^2_\Delta)$. In this case, we can simply group observation patterns based on the number of missing covariates, which naturally partitions the 100 patterns into 10 groups according to d_j. When the target number of groups $\tilde{m}$ is smaller than 10, additional clustering can be applied to merge groups with similar numbers of missing covariates.

A.2 Panel Data with Time Series Dimension

We extend our framework to panel data settings where observations are collected repeatedly over time for each unit. More specifically, consider a panel with observations $Z_{it} = (Y_{it}, X_{it})$ for units $i = 1, \ldots, N$ and time periods $t = 1, \ldots, T$, where $Y_{it} \in \mathbb{R}$ is a scalar outcome and $X_{it} \in \mathbb{R}^L$ is an L-dimensional vector of covariates. We assume that the unit-level time series (Y_i, X_i) are i.i.d. across unit i, while allowing for arbitrary temporal dependence within each unit. In addition, we assume that the marginal distribution of (Y_{it}, X_{it}) is identical across time periods. Under this assumption, the target parameter θ_0 is defined as the unique solution to the population moment condition $G(\theta) = \mathbb{E}[g_\theta(X_{it}, Y_{it})] = 0$.

The presence of a time-series dimension provides additional structure that can be exploited to improve imputation quality. When covariates exhibit persistence over time, imputation functions can leverage both temporal and cross-covariate information to recover missing entries more accurately. For example, one may model a covariate using its lagged value and the contemporaneous values of other covariates as

$$X_{it,l} = \rho_l X_{i,t-1,l} + \beta_l^\top X_{it,-l} + \epsilon_{it,l},$$

where $X_{it,-l}$ denotes all covariates except l. After estimating the parameters using training data, the resulting imputation function takes the form $f_{t,l}^j(X_i, W^{(j)}) = \hat{\rho}_l X_{i,t-1,l} + \hat{\beta}_l^\top X_{it,-l}$ for observations (i,t) with pattern $W^{(j)}$ under which the required covariates are observed.

The weighted debiased moment estimator (5) naturally extends to this panel structure. Incorporating both debiasing and adaptive weighting, the estimator takes the form

$$\begin{aligned} G_N^\lambda(\theta) = \; & \frac{1}{|S_0|} \sum_{(i,t)\in S_0} g_\theta(X_{it}, Y_{it}) + \sum_{k=1}^{K} \sum_{j=1}^{m} \lambda_j \frac{|S_j|}{NT} \Bigg(\frac{1}{|S_j|} \sum_{(i,t)\in S_{j,k}} g_\theta(f_{t,-k}^j(X_i, W^{(j)}), Y_{it}) \\ & - \frac{1}{|S_{0,j}|} \sum_{(i,t)\in S_{0,j,k}} g_\theta(f_{t,-k}^j(X_i, W^{(j)}), Y_{it}) \Bigg). \end{aligned} \tag{A.1}$$

Here, $S_0 = \{(i,t) : X_{it} \text{ are observed}\}$ denotes the set of fully observed data. When using cross-fitting to train the imputation functions, we partition the data into K folds at the unit level to ensure independence between training and estimation samples. For a given observation pattern $W^{(j)}$, not all fully observed (i,t) pairs are valid for debiasing. For example, a unit may be observed at time t but lack lagged values at time $t-1$, in which case it cannot be used to debias imputation error for lag-based models. We therefore construct pattern-specific debiasing sets $S_{0,j} \subseteq S_0$ for each

observation pattern $W^{(j)}$.

When the debiasing sets $S_{0,j}$ differ across observation patterns, a fully observed (i,t) pair may enter a different number of subtraction terms in the estimator (A.1). As a result, some fully observed observations may receive greater aggregate weight than others. To correct for this imbalance, we introduce a normalization adjustment that re-centers the contribution of the fully observed data across all patterns. Specifically, the following terms offsets the additional weight induced by the pattern-specific debiasing terms:

$$\sum_{j=1}^{m} \lambda_j \frac{|S_j|}{NT} \left[\frac{1}{|S_{0,j}|} \sum_{(i,t)\in S_{0,j}} g_\theta(X_{it}, Y_{it}) - \frac{1}{|S_0|} \sum_{(i,t)\in S_0} g_\theta(X_{it}, Y_{it}) \right].$$

This normalization ensures that, after accounting for all debiasing adjustments, each fully observed observation contributes the same aggregate weight to the estimator. In the empirical study in Section 8, we incorporate this normalization term when constructing the estimator.

Finally, we note that the framework can also accommodate time-varying marginal distributions. In this case, we assume that the unit-level time series $Z_i = (Y_i, X_i)$ are i.i.d. draws from some distribution P on $(\mathcal{Y} \times \mathcal{X})^T$. The target parameter can then be defined through the time-averaged population moment

$$G(\theta) = \mathbb{E}_P \left[\frac{1}{T} \sum_{t=1}^{T} g_\theta(X_{it}, Y_{it}) \right] = 0.$$

Under this formulation, our framework extends by treating the entire within-unit time series as a single observation. This simplification comes at the cost that the effective sample size is determined by the cross-sectional dimension N, rather than the total number of observations NT.

A.3 Bias-Variance Analysis of Reweighting

We illustrate the bias-variance tradeoff in the likelihood reweighting of Section 5 with a toy example. Consider a simplified setting with two covariates where X_1 is sometimes missing, while X_2 is always observed. We consider endogenous missingness where the distribution of the covariates differs on the missing data from the overall sample. In more detail, units in the fully observed subset S_0 are drawn i.i.d. from distribution P^0, while units in the partially observed subset S_1 are drawn i.i.d. from distribution P^1, with relative sizes $r_0 = N_0/N$ and $r_1 = N_1/N$.

We assume that X_2 is binary taking values a or b, and $X_1 = X_2 + \Delta$, where $\Delta \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$.

The outcome is generated as

$$Y = I(X_2 = a) \cdot (\beta_{a_1} X_1 + \beta_{a_2} X_2) + I(X_2 = b) \cdot (\beta_{b_1} X_1 + \beta_{b_2} X_2) + \epsilon,$$

where $\epsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. The distributions P^0 and P^1 differ in their marginal distribution of X_2:

$$\begin{aligned} \mathbb{P}_0(X_2 = a) = p_0, \quad \mathbb{P}_0(X_2 = b) = 1 - p_0, \\ \mathbb{P}_1(X_2 = a) = p_1, \quad \mathbb{P}_1(X_2 = b) = 1 - p_1. \end{aligned}$$

This yields likelihood ratios $\omega(a) = p_1/p_0$ and $\omega(b) = (1 - p_1)/(1 - p_0)$. For missing X_1, we use the imputation function $f(X)_1 = X_2$. Suppose that our goal is to estimate the regression coefficients θ_0 that solve

$$G(\theta) = r_0 \mathbb{E}_{P^0}[XX^\top \theta - XY] + r_1 \mathbb{E}_{P^1}[XX^\top \theta - XY] = 0.$$

We now compare the theoretical properties of two approaches for estimating θ_0: our original estimator $\hat{\theta}^\lambda$ that solves $G_N^\lambda(\theta) = 0$ with G_N^λ defined in (5), and the reweighted estimator $\hat{\theta}^{\lambda,\omega}$ solving $G_N^{\lambda,\omega}(\theta) = 0$ with the likelihood reweighted moment defined in (11).

As established in Theorem 3, $\hat{\theta}^{\lambda,\omega}$ is asymptotically unbiased and asymptotically normal:

$$\sqrt{N_0}(\hat{\theta}^{\lambda,\omega} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, J_{\theta_0}^{-1} \left(\Sigma_{\Delta,0}^\lambda + \lambda^2 r_1 r_0 \Sigma_{f,1}\right) J_{\theta_0}^{-1}\right),$$

where $J_\theta = \nabla_\theta G(\theta)$, $\Sigma_{f,1} = \mathrm{Cov}_{P^1}(g_{\theta_0}(f(X), Y))$, and

$$\Sigma_{\Delta,0}^\lambda = \mathrm{Cov}_{P^0}\left(r_0 g_{\theta_0}(X, Y) + r_1 g_{\theta_0}(X, Y)\omega(X_2) - \lambda r_1 g_{\theta_0}(f(X), Y)\omega(X_2)\right).$$

In contrast, while the original estimator $\hat{\theta}^\lambda$ is also asymptotically normal, it has an asymptotic bias equal to $\tilde{\theta}^\lambda - \theta_0$:

$$\sqrt{N_0}(\hat{\theta}^\lambda - \tilde{\theta}^\lambda) \xrightarrow{d} \mathcal{N}\left(0, \tilde{J}_{\tilde{\theta}}^{-1} \left(\tilde{\Sigma}_{\Delta,0}^\lambda + \lambda^2 r_1 r_0 \tilde{\Sigma}_{f,1}\right) \tilde{J}_{\tilde{\theta}}^{-1}\right),$$

where $\tilde{\theta}^\lambda$ solves the moment condition that depends on the distributional shift $p_0 - p_1$:

$$\tilde{G}^\lambda(\theta) = G(\theta) + (1 - \lambda) r_1 (p_0 - p_1)(h_\theta(a) - h_\theta(b)) + \lambda r_1 \mu (p_0 - p_1)(g_\theta(a) - g_\theta(b)) = 0.$$

The conditional moments that determine the bias are

$$h_\theta(a) = \mathbb{E}[XX^\top\theta - XY \mid X_2 = a] = \begin{bmatrix} (\theta_1 - \beta_{a_1})(\sigma^2 - (\mu+a)^2) + (\theta_2 - \beta_{a_2})a(\mu+a) \\ a(\theta_1 - \beta_{a_1})(\mu+a) + (\theta_2 - \beta_{a_2})a^2 \end{bmatrix},$$

and

$$g_\theta(a) = \mathbb{E}\left[\begin{bmatrix} 2X_1\theta_1 + X_2\theta_2 - Y \\ \theta_1 X_2 \end{bmatrix} \,\middle|\, X_2 = a \right] = \begin{bmatrix} (2\theta_1 - \beta_{a_1})(\mu+a) + (\theta_2 - \beta_{a_2})a \\ a\theta_1 \end{bmatrix},$$

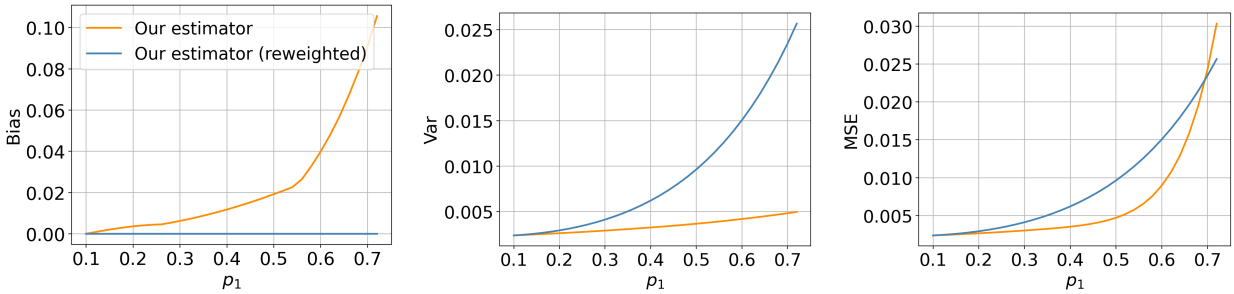
with $h_\theta(b)$ and $g_\theta(b)$ defined analogously. The asymptotic variance depends on $\tilde{J}_\theta = \nabla\tilde{G}(\theta)$, $\Sigma_{f,1} = \mathrm{Cov}_{P^1}(g_{\tilde{\theta}}(f(X), Y))$, and

$$\Sigma^\lambda_{\Delta,0} = \mathrm{Cov}_{P^0}\left(g_{\theta_0}(X,Y) - \lambda r_1 g_{\theta_0}(f(X),Y)\right).$$

Figure A.1 illustrates the theoretical bias and standard deviation of the two estimators. We fix $p_0 = 0.1$ and vary p_1 from 0.1 to 0.7 to examine how the distributional shift affects bias and variance.

As demonstrated in Figure A.1, there is a bias-variance tradeoff with likelihood reweighting. As the distributional shift $|p_1 - p_0|$ increases, the bias of the original estimator increases, while the reweighted estimator remains unbiased throughout. However, the variance of the reweighted estimator increases substantially with the distributional shift. The original estimator maintains a stable variance regardless of p_1. Neither method dominates in terms of mean squared error. The optimal choice depends on multiple elements, including the degree of distributional heterogeneity, the signal-to-noise ratio in the outcome, and the imputation accuracy.

Figure A.1: Bias-variance tradeoff in likelihood reweighting



(a) Bias **(b)** Var **(c)** MSE

This figure compares the theoretical bias and variance of our original estimator (without reweighting) and the reweighted variant. Panel (a) shows average absolute bias, panel (b) shows total variance, and panel (c) shows mean squared error (MSE) as functions of p_1, where $p_0 = 0.1$ is fixed. We use the parameters $N = 10000$, $N_0 = 1000$, $a = 1$, $b = 2$, $\beta_{a1} = \beta_{a2} = 1$, $\beta_{b1} = \beta_{b2} = 2$, $\mu = 0.3$, $\sigma = 1$, and $\sigma_\epsilon = 5$. Subset weights λ are optimized to minimize asymptotic variance for each estimator.

A.4 Additional Simulation Results

Table A.1: Comparison of different estimators (# observation patterns = 20 and $\sigma_\epsilon = 2$)

		$k_0 = 5, \hat{k} = 5$				$k_0 = 5, \hat{k} = 3$			
p	Method	$\bar{\lambda}$	|Bias|	Var	MSE	$\bar{\lambda}$	|Bias|	Var	MSE
40%	**Our method**	0.52	**0.006**	**0.102**	**0.104**	0.31	**0.005**	**0.111**	**0.112**
	Complete subset	0.0	**0.004**	0.130	0.131	0.0	**0.004**	0.130	0.131
	Naive imputation	-	0.041	0.339	0.375	-	0.164	3.122	3.709
60%	**Our method**	0.59	**0.006**	**0.094**	**0.095**	0.35	**0.005**	**0.105**	**0.106**
	Complete subset	0.0	**0.004**	0.130	0.131	0.0	**0.004**	0.130	0.131
	Naive imputation	-	0.020	0.129	0.139	-	0.153	2.238	2.764
80%	**Our method**	0.71	**0.005**	**0.061**	**0.062**	0.53	**0.005**	**0.075**	**0.075**
	Complete subset	0.0	**0.004**	0.130	0.131	0.0	**0.004**	0.130	0.131
	Naive imputation	-	0.011	0.066	0.069	-	0.117	1.140	1.466

This table reports the average absolute bias (|Bias|), total variance (Var), and mean squared error (MSE) for three methods: our method, complete-case analysis (complete subset), and naive use of imputation (naive imputation). For our method, we use $K = 10$ folds for cross-fitting and $B = 50$ bootstrap samples for weight selection. Bold numbers indicate the (near) best performance across models. The data-generating process and covariate observation patterns are as follows: $\sigma_e = 1$, $\sigma_\epsilon = 2$, and the true regression coefficient is $\theta_0 = \mathbf{1}_L$ (vector of ones). We assume 19 partially observed patterns where the first half contains 60% observed covariates and the second half contains p observed covariates with $p \in \{40\%, 60\%, 80\%\}$. Each partially observed unit is randomly assigned to one of the 19 partially observed patterns. We set $N = 5000$, $N_0 = 500$, $L = 20$, and run 200 simulations for each setup.

Table A.2: Comparison of different estimators (# observation patterns = 10 and $\sigma_\epsilon = 3$)

		$k_0 = 5, \hat{k} = 5$				$k_0 = 8, \hat{k} = 5$			
p	Method	$\bar{\lambda}$	|Bias|	Var	MSE	$\bar{\lambda}$	|Bias|	Var	MSE
40%	**Our method**	0.70	**0.008**	**0.204**	**0.206**	0.45	**0.008**	**0.225**	**0.227**
	Complete subset	0.0	**0.007**	0.293	0.295	0.0	**0.007**	0.293	0.295
	Naive imputation	-	0.040	0.418	0.458	-	0.146	3.858	4.391
60%	**Our method**	0.84	**0.009**	0.180	**0.182**	0.51	**0.007**	**0.207**	**0.209**
	Complete subset	0.0	**0.007**	0.293	0.295	0.0	**0.007**	0.293	0.295
	Naive imputation	-	0.020	**0.176**	0.186	-	0.125	2.588	2.980
80%	**Our method**	0.86	**0.007**	0.126	0.128	0.63	**0.006**	**0.156**	**0.157**
	Complete subset	0.0	**0.007**	0.293	0.295	0.0	**0.007**	0.293	0.295
	Naive imputation	-	0.013	**0.100**	**0.105**	-	0.103	1.522	1.813

This table reports the average absolute bias (|Bias|), total variance (Var), and mean squared error (MSE) for three methods: our method, complete-case analysis (complete subset), and naive use of imputation (naive imputation). For our method, we use $K = 10$ folds for cross-fitting and $B = 50$ bootstrap samples for weight selection. Bold numbers indicate the (near) best performance across models. The data-generating process and covariate observation patterns are as follows: $\sigma_e = 1$, $\sigma_\epsilon = 3$, and the true regression coefficient is $\theta_0 = \mathbf{1}_L$ (vector of ones). We assume 9 partially observed patterns where the first half contains 60% observed covariates and the second half contains p observed covariates with $p \in \{40\%, 60\%, 80\%\}$. Each partially observed unit is randomly assigned to one of the 9 partially observed patterns. We set $N = 5000$, $N_0 = 500$, $L = 20$, and run 200 simulations for each setup.

Table A.3: Comparison of different estimators (# observation patterns = 10 and $\sigma_\epsilon = 2$)

		$k_0 = 5, \hat{k} = 5$				$k_0 = 8, \hat{k} = 5$			
p	Method	$\bar{\lambda}$	|Bias|	Var	MSE	$\bar{\lambda}$	|Bias|	Var	MSE
40%	**Our method**	0.46	**0.005**	**0.104**	**0.105**	0.26	**0.005**	**0.112**	**0.113**
	Complete subset	0.0	**0.004**	0.130	0.131	0.0	**0.004**	0.130	0.131
	Naive imputation	-	0.040	0.374	0.412	-	0.146	3.835	4.365
60%	**Our method**	0.53	**0.006**	**0.097**	**0.098**	0.31	**0.005**	**0.106**	**0.107**
	Complete subset	0.0	**0.004**	0.130	0.131	0.0	**0.004**	0.130	0.131
	Naive imputation	-	0.019	0.142	0.152	-	0.125	2.567	2.958
80%	**Our method**	0.65	**0.005**	**0.070**	**0.071**	0.47	**0.004**	**0.084**	**0.085**
	Complete subset	0.0	**0.004**	0.130	0.131	0.0	**0.004**	0.130	0.131
	Naive imputation	-	0.013	0.076	0.081	-	0.103	1.502	1.793

This table reports the average absolute bias (|Bias|), total variance (Var), and mean squared error (MSE) for three methods: our method, complete-case analysis (complete subset), and naive use of imputation (naive imputation). For our method, we use $K = 10$ folds for cross-fitting and $B = 50$ bootstrap samples for weight selection. Bold numbers indicate the (near) best performance across models. The data-generating process and covariate observation patterns are as follows: $\sigma_e = 1$, $\sigma_\epsilon = 2$, and the true regression coefficient is $\theta_0 = \mathbf{1}_L$ (vector of ones). We assume 9 partially observed patterns where the first half contains 60% observed covariates and the second half contains p observed covariates with $p \in \{40\%, 60\%, 80\%\}$. Each partially observed unit is randomly assigned to one of the 9 partially observed patterns. We set $N = 5000$, $N_0 = 500$, $L = 20$, and run 200 simulations for each setup.

A.5 Empirical Study

A.5.1 Definitions and Summary Statistics of Variables

Table A.4: Variable definitions

Variable	Definition	Data Source
Return	Monthly stock return.	CRSP
Log scope 1 emission	Natural logarithm of scope 1 emissions (measured in tCO_2e). Scope 1 emissions are direct emissions from owned or controlled facilities, including fossil fuel use in production.	Trucost
Log scope 2 emission	Natural logarithm of scope 2 emissions (measured in tCO_2e). Scope 2 emissions come from purchased heat, steam, and electricity consumed by the company.	Trucost
Log scope 3 emission	Natural logarithm of scope 3 emissions (measured in tCO_2e). Scope 3 emissions are caused by company operations and products but occur from sources not owned or controlled by the company.	Trucost
Firm size	Natural logarithm of firm's total market capitalization.	Compustat
Book-to-market	Book value of equity divided by market value of equity.	Compustat
ROE	Net income divided by the value of equity.	Compustat
I/A	Capital expenditures divided by year-end total assets.	Compustat
LogPPE	Natural logarithm of property, plant, and equipment.	Compustat
Leverage	Long-term debt divided by assets.	Compustat
SalesGR	Year-over-year change in annual firm revenues, normalized by prior-year revenue.	Compustat
EPSGR	Year-over-year change in earnings per share, normalized by prior-year EPS.	Compustat
Total assets	Total assets of a firm in a given year.	Compustat
Log sales	Natural logarithm of total sales in a given year.	Compustat
HHI	Herfindahl concentration index of sector-level firm sales relative to the full firm (within-firm measure).	Compustat
Volatility	Monthly stock return volatility calculated over the past year.	CRSP
Momentum	Total stock return over the past 12 months, excluding the most recent month.	CRSP
Beta	CAPM beta calculated over the past year.	CRSP

This table provides definitions and data sources of the variables used in our empirical study. The variables from CRSP and Compustat are directly downloaded from the CRSP–Compustat merged database accessed through WRDS.

Table A.5: Summary statistics

Variable	Winsorized	Mean	Std	Median
Return	–	1.186	15.575	0.809
Log scope 1 emission	–	11.752	2.911	11.570
Log scope 2 emission	–	11.643	1.988	11.709
Log scope 3 emission	–	14.117	1.420	14.211
Firm size	–	8.048	1.731	8.064
Book-to-market	2.5	0.416	0.350	0.335
ROE	2.5	4.267	41.374	10.532
I/A	2.5	0.041	0.040	0.028
LogPPE	–	6.665	2.214	6.788
Leverage	2.5	0.244	0.194	0.224
SalesGR	0.5	0.132	0.420	0.052
EPSGR	0.5	0.002	2.133	0.000
HHI	–	0.350	0.236	0.333
Volatility	0.5	0.115	0.071	0.097
Momentum	0.5	0.142	0.518	0.081
Beta	–	1.164	1.324	1.090

This table reports summary statistics and winsorization levels for variables used in our regression analysis. Winsorization levels indicate the percentile cutoffs applied to each tail of the distribution, following Bolton and Kacperczyk (2021) and Aswani, Raghunandan, and Rajgopal (2024). For emissions data, summary statistics are calculated using only actually reported (non-imputed) observations.

A.5.2 Removal of Fixed Effects

To implement the regression in equation (14), we first apply a two-way within transformation to remove the time and industry fixed effects. For notational simplicity, let $Y_{it} = RET_{it}$ denote the monthly stock return and let $X_{it} = [LOG(Emission)_{it}, Controls_{it}]$ denote the covariate vector for firm i at time t. The regression model can be then written as

$$Y_{it} = \beta^\top X_{it} + \alpha + \delta_{g(i)} + \gamma_t + \epsilon_{it}.$$

We remove the fixed effects through a two-way within transformation. For any variable Z_{it}, define $\bar{Z}_{g(i)\cdot}$ as the mean of Z_{it} within industry $g(i)$, $\bar{Z}_{\cdot t}$ as the mean within time period t, and $\bar{Z}$ as the overall mean. The transformed variable is given by

$$\tilde{Z}_{it} = Z_{it} - \bar{Z}_{g(i)\cdot} - \bar{Z}_{\cdot t} + \bar{Z}.$$

Applying this transformation to both the outcome and covariates yields an approximate representation of the regression (14) in terms of the transformed variables

$$\tilde{Y}_{it} \approx \beta^\top \tilde{X}_{it} + \tilde{\epsilon}_{it},$$

which we use to remove the fixed effects in the unbalanced panel with missing covariates. In the complete-case analysis, all quantities involved in the transformation are computed using only the fully observed subset, whereas in the other approaches they are computed using all available observations. The transformed variables $\tilde{Y}_{it}$ and $\tilde{X}_{it}$ are used in the estimation procedure of all approaches.

Imputation of the within-transformed covariates corresponds to explaining variation in the covariates that is not captured by time and industry fixed effects. Duan, Pelger, and Xiong (2024a) discuss the importance of applying within-transformation for imputation in the context of latent factor models.

A.5.3 Imputation Methods

This section describes the imputation methods used in the empirical study. For each emission scope and each partially observed subset, we consider three classes of imputation methods: ridge regression, random forest, and factor model imputation. Ridge regression and random forests require observations with reported emissions for training, whereas the factor model approach can additionally leverage partially observed data through its latent factor structure. All methods are evaluated using fully observed data.

Let E_{it} denote log emissions for firm i at time t, and let Z_{it} denote the vector of observed covariates used for imputation. For subset S_1, Z_{it} includes both cross-sectional firm characteristics and within-firm temporal emissions from other periods; for subset S_2, Z_{it} includes only cross-sectional firm characteristics.

Ridge regression. Ridge regression imputes missing emissions by modeling emissions as a linear function of the observed covariates. The ridge estimator is defined as

$$\hat{\beta}_{\text{Ridge}} = \arg\min_{\beta} \sum_{(i,t)\in S_{\text{train}}} \left(E_{it} - Z_{it}^\top \beta\right)^2 + \lambda_{\text{Ridge}} \|\beta\|_2^2,$$

where S_{train} denotes the training set of firm-month observations with reported emissions and λ_{Ridge} is a regularization parameter. Missing emissions are then imputed as

$$\widehat{E}_{it}^{\text{Ridge}} = Z_{it}^{\top}\hat{\beta}_{\text{Ridge}}.$$

The regularization parameter λ_{Ridge} is selected via five-fold cross-validation to maximize out-of-sample R^2. We perform a grid search over a logarithmically spaced grid of λ_{Ridge} values ranging from 10^{-6} to 10^{6}.

Random forest. Random forest provides a flexible nonlinear imputation method that captures interactions and nonlinearities among covariates. The random forest imputation is

$$\widehat{E}_{it}^{\text{RF}} = \frac{1}{B}\sum_{b=1}^{B} T_b(Z_{it}),$$

where $\{T_b\}_{b=1}^{B}$ are regression trees trained on bootstrap samples of the training set of firm-month observations with reported emissions. Each tree is constructed by recursively partitioning the covariate space to minimize squared prediction error within terminal nodes, with random subsampling of covariates at each split. We implement random forests using the standard algorithm as implemented in scikit-learn (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, et al., 2011). Hyperparameters, including the number of features considered at each split (max_features), maximum tree depth (max_depth), minimum leaf size (min_samples_leaf), and subsample size (max_samples), are selected via cross-validation to maximize out-of-sample R^2. Specifically, we search over max_features $\in$ {"sqrt", 0.5, 0.8}, max_depth $\in$ {10, 20}, min_samples_leaf $\in$ {2, 3, 5}, and max_samples $\in$ {0.8, 0.9}.

Factor model imputation. Factor model imputation follows the procedure described in Section 3.3, with two modifications tailored to the empirical setting. First, we add a ridge regularization term to the regression of observed characteristics on the estimated factors $\hat{F}$:

$$\widehat{E}_{it}^{\text{FM}} = \hat{F}_l^{\top}\left(\sum_{h=1}^{L} W_h^{(j)}\hat{F}_h\hat{F}_h^{\top} + \lambda I_k\right)^{-1}\left(\sum_{h=1}^{L} W_h^{(j)}\hat{F}_h X_h\right),$$

where λ is a ridge penalty parameter. Bryzgalova, Lerner, Lettau, and Pelger (2025) shows that the finite sample performance of the latent factor model can be improved through ridge regularization.

Second, for observations in subset S_1 with available temporal emissions, we further refine the imputed values by regressing observed emissions on both the factor-model imputed values and the available temporal emissions, and use the fitted value from this regression as the final imputed emission. The number of factors and the ridge parameter λ are selected via cross-validation to maximize out-of-sample R^2. Specifically, we search over the set $\{1, 2, 3, 4, 5, 8, 10\}$ for the number of factors and over a logarithmically spaced grid of λ values ranging from 10^{-6} to 10^{6}.

A.6 Proof

A.6.1 Proof of Proposition 1

The result follows from Theorem 2, which derives the asymptotic distribution of the debiased estimator. We start with calculating the matrix $\Sigma_{f,1}$ in the asymptotic variance. Let $\tilde{X}_i = f^1(X_i)$ denote the imputed vector for unit i under the partially observed pattern. The (k,l)-th entry of $\Sigma_{f,1}$ is given by

$$\begin{aligned}(\Sigma_{f,1})_{kl} = &\ \mathrm{Cov}\left(\tilde{X}_{ik}\left(\sum_{q=L_0+1}^{L}\Delta_{iq}\theta_{0,q}-\epsilon_i\right),\tilde{X}_{il}\left(\sum_{q=L_0+1}^{L}\Delta_{iq}\theta_{0,q}-\epsilon_i\right)\right)\\ = &\sum_{q,r=L_0+1}^{L}\theta_{0,r}\theta_{0,q}\left[\mathrm{Cov}(X_{ik}\Delta_{ir},X_{il}\Delta_{iq})+\mathrm{Cov}(\Delta_{ik}\Delta_{ir},\Delta_{il}\Delta_{iq})\right]+\mathrm{Cov}(X_{ik}\epsilon_i,X_{il}\epsilon_i)\\ &+\mathrm{Cov}(\Delta_{ik}\epsilon_i,\Delta_{il}\epsilon_i).\end{aligned}$$

When $k \le L_0$ or $l \le L_0$, we have

$$(\Sigma_{f,1})_{kl} = \left(\sigma_\epsilon^2+\theta_0^{(2)}\sigma_\Delta^2+(\theta_0^{(1)})^2\mu_\Delta^2\right)\cdot\Sigma_X,$$

where $\theta_0^{(1)} = \sum_{l=L_0+1}^{L}\theta_{0,l}$ and $\theta_0^{(2)} = \sum_{l=L_0+1}^{L}\theta_{0,l}^2$. When $k=l>L_0$, we obtain

$$\begin{aligned}(\Sigma_{f,1})_{kk} = &\sum_{q,r=L_0+1}^{L}\theta_{0,r}\theta_{0,q}\left((\Sigma_X)_{kk}\mathbb{E}[\Delta_{ir}\Delta_{iq}]+\mathbb{E}[\Delta_{ik}^2\Delta_{ir}\Delta_{iq}]-\mathbb{E}[\Delta_{ik}\Delta_{ir}]\cdot\mathbb{E}[\Delta_{ik}\Delta_{iq}]\right)\\ &+\sigma_\epsilon^2(\Sigma_X)_{kk}+\sigma_\epsilon^2\sigma_\Delta^2\\ = &\left(\sigma_\epsilon^2+\theta_0^{(2)}\sigma_\Delta^2+(\theta_0^{(1)})^2\mu_\Delta^2\right)(\Sigma_X)_{kk}+(\sigma_\Delta^2+\mu_\Delta^2)\sigma_\epsilon^2+(\theta_0^{(2)}+\theta_{0k}^2)\sigma_\Delta^4\\ &+\left((\theta_0^{(1)})^2+2\theta_{0k}\theta_0^{(1)}+\theta_0^{(2)}\right)\mu_\Delta^2\sigma_\Delta^2.\end{aligned}$$

When $k,l>L_0$ and $k\neq l$,

$$\begin{aligned}(\Sigma_{f,1})_{kl} = &\left(\sigma_\epsilon^2+\theta_0^{(2)}\sigma_\Delta^2+(\theta_0^{(1)})^2\mu_\Delta^2\right)(\Sigma_X)_{kl}+\mu_\Delta^2\sigma_\epsilon^2+\theta_{0,k}\theta_{0,l}\sigma_\Delta^4\\ &+\left((\theta_{0k}+\theta_{0l})\theta_0^{(1)}+\theta_0^{(2)}\right)\mu_\Delta^2\sigma_\Delta^2.\end{aligned}$$

Therefore, we have $\Sigma_{f,1} = (\sigma_\epsilon^2+\theta_0^{(2)}\sigma_\Delta^2+(\theta_0^{(1)})^2\mu_\Delta^2)\Sigma_X+\Sigma_1$, where $(\Sigma_1)_{kl} = I(k,l>L_0)(\mu_\Delta^2\sigma_\epsilon^2+\theta_{0k}\theta_{0l}\sigma_\Delta^4+((\theta_{0k}+\theta_{0l})\theta_0^{(1)}+\theta_0^{(2)})\mu_\Delta^2\sigma_\Delta^2)+I(k=l>L_0)(\sigma_\epsilon^2\sigma_\Delta^2+\theta_0^{(2)}\sigma_\Delta^4+(\theta_0^{(1)})^2\mu_\Delta^2\sigma_\Delta^2)$.

Next, we calculate the matrix $\Sigma^{\lambda}_{\Delta,0}$, which is given by

$$\begin{aligned}\Sigma^{\lambda}_{\Delta,0} = &\ \mathrm{Cov}(g_{\theta_0}(X_i, Y_i) - \lambda r_1 g_{\theta_0}(\tilde{X}_i, Y_i)) \\ = &\ \mathrm{Cov}(g_{\theta_0}(X_i, Y_i)) - 2r_1\lambda \mathrm{Cov}(g_{\theta_0}(X_i, Y_i), g_{\theta_0}(\tilde{X}_i, Y_i)) + r_1^2\lambda^2\Sigma_{f,1}.\end{aligned}$$

Since $\mathrm{Cov}(g_{\theta_0}(X_i, Y_i)) = \mathrm{Cov}(g_{\theta_0}(X_i, Y_i), g_{\theta_0}(\tilde{X}_i, Y_i)) = \sigma_\epsilon^2\Sigma_X$, we get

$$\Sigma^{\lambda}_{\Delta,0} = (1 - 2r_1\lambda)\sigma_\epsilon^2\Sigma_X + r_1^2\lambda^2\Sigma_{f,1}.$$

Using $J_{\theta_0} = \mathbb{E}[X_i X_i^\top] = \Sigma_X$, the asymptotic covariance of $\hat{\theta}^{\lambda}$ becomes

$$\Sigma^{\lambda} = \left[(1 - 2r_1\lambda)\sigma_\epsilon^2 + \lambda^2 r_1 \left(\sigma_\epsilon^2 + \theta_0^{(2)}\sigma_\Delta^2 + (\theta_0^{(1)})^2\mu_\Delta^2\right)\right]\Sigma_X^{-1} + \lambda^2 r_1 \Sigma_X^{-1}\Sigma_1\Sigma_X^{-1}.$$

The average asymptotic variance $\mathrm{tr}(\Sigma^{\lambda})$ is given by

$$\mathrm{tr}(\Sigma^{\lambda}) = \ \left[(1 - 2r_1\lambda)\sigma_\epsilon^2 + \lambda^2 r_1 \left(\sigma_\epsilon^2 + \theta_0^{(2)}\sigma_\Delta^2 + (\theta_0^{(1)})^2\mu_\Delta^2\right)\right]\mathrm{tr}(\Sigma_X^{-1}) + \lambda^2 r_1 \mathrm{tr}(\Sigma_X^{-1}\Sigma_1\Sigma_X^{-1}).$$

Minimizing $\mathrm{tr}(\Sigma^{\lambda})$ yields

$$\lambda^* = \frac{\sigma_\epsilon^2 \cdot \mathrm{tr}(\Sigma_X^{-1})}{(\sigma_\epsilon^2 + \theta_0^{(2)}\sigma_\Delta^2 + (\theta_0^{(1)})^2\mu_\Delta^2)\,\mathrm{tr}(\Sigma_X^{-1}) + \mathrm{tr}(\Sigma_X^{-1}\Sigma_1\Sigma_X^{-1})}.$$

A.6.2 Proof of Theorem 1

We first establish the consistency of the debiased moment estimator $G_N^{\lambda}(\theta)$ in (5).

Lemma A.1. *Under Assumptions 1 and 2.1, for any λ and $\theta \in \Theta$, as $N_0 \to \infty$:*

$$G_N^{\lambda}(\theta) - G(\theta) = o_p(1).$$

Proof. Observe that the difference between the empirical moment function $G_N^{\lambda}(\theta)$ and its population

counterpart $G(\theta)$ can be decomposed as:

$$\begin{aligned} G_N^{\lambda}(\theta) - G(\theta) = \; & \frac{1}{N_0} \sum_{i \in S_0} \left(g_\theta(X_i, Y_i) - \mathbb{E}[g_\theta(X_i, Y_i)] \right) \\ & + \sum_{k=1}^{K} \sum_{j=1}^{m} \lambda_j \frac{N_j}{N} \Bigg[\frac{1}{N_j} \sum_{i \in S_{j,k}} \Big(g_\theta(f_{-k}^j(X_i), Y_i) - \mathbb{E}_Z[g_\theta(f_{-k}^j(X_i), Y_i)] \Big) \\ & - \frac{1}{N_0} \sum_{i \in S_{0,k}} \Big(g_\theta(f_{-k}^j(X_i), Y_i) - \mathbb{E}_Z[g_\theta(f_{-k}^j(X_i), Y_i)] \Big) \Bigg], \end{aligned}$$

where $\mathbb{E}_Z[\cdot]$ denotes the expectation taken with respect to the randomness in $Z_i = (X_i, Y_i)$, treating the learned imputation function $f_{-k}^j(\cdot) = f^j(\cdot; [N] \backslash \tilde{S}_k)$ as fixed (conditional on the training sample $[N] \backslash \tilde{S}_k$). By Assumption 2.1, all moment functions have finite variance. The law of large number yields

$$\frac{1}{N_0} \sum_{i \in S_0} \left(g_\theta(X_i, Y_i) - \mathbb{E}[g_\theta(X_i, Y_i)] \right) = o_p(1),$$

and for each j with $N_j \to \infty$,

$$\frac{1}{N_j} \sum_{i \in S_{j,k}} \Big(g_\theta(f_{-k}^j(X_i), Y_i) - \mathbb{E}_Z[g_\theta(f_{-k}^j(X_i), Y_i)] \Big) = o_p(1).$$

This completes our proof. □

Lemma A.2. *Define $B_\epsilon(\theta_0) = \{\theta \in \Theta : \|\theta - \theta_0\| \le \epsilon\}$. Under Assumptions 1, 2.1 and 2.3. For any λ and fixed $\varepsilon > 0$ sufficiently small,*

$$\sup_{\theta \in B_\varepsilon(\theta_0)} \left\| G_N^{\lambda}(\theta) - G(\theta) \right\| = o_p(1).$$

Proof. For any fixed $\varepsilon, \delta > 0$, let $\mathcal{N}_\delta \subset B_\epsilon(\theta_0)$ be a finite δ-net of $B_\epsilon(\theta_0)$, i.e., for every $\theta \in B_\epsilon(\theta_0)$ there exists $\theta' \in \mathcal{N}_\delta$ such that $\|\theta - \theta'\| \le \delta$. Since $B_\epsilon(\theta_0)$ is compact, such a finite net exists.

By Lemma A.1, for each fixed $\theta \in \mathcal{N}_\delta$, $G_N^{\lambda}(\theta) - G(\theta) = o_p(1)$. Because $\mathcal{N}_\delta$ is finite, a union bound yields

$$\max_{\theta \in \mathcal{N}_\delta} \|G_N^{\lambda}(\theta) - G(\theta)\| = o_p(1).$$

By Assumption 2.3, for any $\theta, \theta' \in B_\varepsilon(\theta_0)$,

$$\|G_N^{\lambda}(\theta) - G_N^{\lambda}(\theta')\| \le C_1 \|\theta - \theta'\| \text{ with } C_1 = O_p(1),$$

and $\|G(\theta) - G(\theta')\| \leq C_2\|\theta - \theta'\|$ with $C_2 < \infty$.

Take any $\theta \in B_\varepsilon(\theta_0)$ and choose $\theta' \in \mathcal{N}_\delta$ with $\|\theta - \theta'\| \leq \delta$. Then by the triangle inequality,

$$\|G_N^\lambda(\theta) - G(\theta)\| \leq \|G_N^\lambda(\theta) - G_N^\lambda(\theta')\| + \|G_N^\lambda(\theta') - G(\theta')\| + \|G(\theta') - G(\theta)\|.$$

Taking the supremum over $\theta \in B_\varepsilon(\theta_0)$ yields

$$\sup_{\theta \in B_\varepsilon(\theta_0)} \|G_N^\lambda(\theta) - G(\theta)\| \leq \max_{\theta \in \mathcal{N}_\delta} \|G_N^\lambda(\theta) - G(\theta)\| + (C_1 + C_2)\delta = (C_1 + C_2)\delta + o_p(1).$$

We let $\delta \to 0$ and get

$$\sup_{\theta \in B_\varepsilon(\theta_0)} \|G_N^\lambda(\theta) - G(\theta)\| = o_p(1),$$

as desired. □

In the following, we consider the two cases stated in Theorem 1 separately.

Case 1. When the parameter space Θ is compact and Assumption 2.3 holds globally on Θ, following similar arguments as in Lemma A.2 we can prove that

$$\sup_{\theta \in \Theta} \big\|G_N^\lambda(\theta) - G(\theta)\big\| = o_p(1),$$

that is, $G_N^\lambda(\theta)$ converges uniformly to $G(\theta)$ on Θ.

Since $G(\theta)$ is continuous, Θ is compact, and $G(\theta) = 0$ admits the unique solution θ_0, it follows that

$$\inf_{\theta \in \Theta:\, \|\theta - \theta_0\| \geq \varepsilon} \|G(\theta)\| > 0.$$

Now, for any $\theta \in \Theta$,

$$\|G_N^\lambda(\theta)\| \;\geq\; \|G(\theta)\| - \|G_N^\lambda(\theta) - G(\theta)\|.$$

For any $\varepsilon > 0$, taking the infimum over $\theta \in \{\theta \in \Theta :\, \|\theta - \theta_0\| \geq \varepsilon\} = A_\varepsilon(\theta_0)$, we obtain

$$\inf_{\theta \in A_\varepsilon(\theta_0)} \|G_N^\lambda(\theta)\| \;\geq\; \inf_{\theta \in A_\varepsilon(\theta_0)} \|G(\theta)\| - \sup_{\theta \in A_\varepsilon(\theta_0)} \|G_N^\lambda(\theta) - G(\theta)\|.$$

Combining this with the uniform convergence of $G_N^\lambda(\theta)$ to $G(\theta)$ implies

$$\mathbb{P}\left(\inf_{\theta \in A_\varepsilon(\theta_0)} \|G_N^\lambda(\theta)\| > 0\right) \to 1.$$

Therefore, with probability tending to one, the equation $G_N^\lambda(\theta) = 0$ admits no solution in $A_\varepsilon(\theta_0)$. Since $\hat{\theta}^\lambda$ is the solution to this equation with probability tending to one, we conclude that

$$\mathbb{P}\big(\|\hat{\theta}^\lambda - \theta_0\| < \varepsilon\big) \to 1.$$

As $\varepsilon > 0$ is arbitrary, this establishes that

$$\hat{\theta}^\lambda \xrightarrow{p} \theta_0.$$

Case 2. In the following, we prove consistency under the alternative assumption that the moment estimator $G_N^\lambda(\theta)$ is nondecreasing with respect to θ, i.e.,

$$(G_N^\lambda(\theta_1) - G_N^\lambda(\theta_2))^\top(\theta_1 - \theta_2) \geq 0, \quad \text{for all } \theta_1, \theta_2 \in \Theta.$$

Define $\partial B_\epsilon(\theta_0) = \{\theta \in \Theta : \|\theta - \theta_0\| = \epsilon\}$. By Assumption 2.2, $G(\theta)$ is differentiable at θ_0 with positive definite Jacobian $J_{\theta_0} = \nabla G(\theta_0)$. Hence there exist constants $c > 0$ and $\varepsilon > 0$ such that, for all $\tilde{\theta} \in \partial B_\varepsilon(\theta_0)$,

$$G(\tilde{\theta})^\top(\tilde{\theta} - \theta_0) = (G(\tilde{\theta}) - G(\theta_0))^\top(\tilde{\theta} - \theta_0) \geq c\varepsilon^2.$$

By Lemma A.2, $G_N^\lambda(\theta)$ converges uniformly to $G(\theta)$ on $B_\varepsilon(\theta_0)$. Therefore, uniformly over $\tilde{\theta} \in \partial B_\varepsilon(\theta_0)$,

$$G_N^\lambda(\tilde{\theta})^\top(\tilde{\theta} - \theta_0) = G(\tilde{\theta})^\top(\tilde{\theta} - \theta_0) + \big(G_N^\lambda(\tilde{\theta}) - G(\tilde{\theta})\big)^\top(\tilde{\theta} - \theta_0) \geq c\varepsilon^2 - o_p(1).$$

For any $\theta \in \Theta$ such that $\|\theta - \theta_0\| \geq 2\varepsilon$, define its projection onto $\partial B_\varepsilon(\theta_0)$ by

$$\tilde{\theta} = \theta_0 + \varepsilon \frac{\theta - \theta_0}{\|\theta - \theta_0\|} \in \partial B_\varepsilon(\theta_0).$$

Since $G_N^\lambda(\cdot)$ is nondecreasing, we have

$$\big(G_N^\lambda(\theta) - G_N^\lambda(\tilde{\theta})\big)^\top(\theta - \tilde{\theta}) \geq 0,$$

which implies $G_N^\lambda(\theta)^\top(\tilde{\theta}-\theta) \leq G_N^\lambda(\tilde{\theta})^\top(\tilde{\theta}-\theta)$. Noting that

$$\tilde{\theta}-\theta_0 = \frac{\varepsilon}{\|\theta-\tilde{\theta}\|}(\theta-\tilde{\theta}),$$

we obtain

$$G_N^\lambda(\tilde{\theta})^\top(\tilde{\theta}-\theta) = -\frac{\|\theta-\tilde{\theta}\|}{\varepsilon} G_N^\lambda(\tilde{\theta})^\top(\tilde{\theta}-\theta_0).$$

Combining this with the uniform bound yields, with probability tending to one,

$$G_N^\lambda(\theta)^\top(\tilde{\theta}-\theta) \leq -\frac{1}{2}c\varepsilon\|\theta-\tilde{\theta}\| \leq -\frac{1}{2}c\varepsilon^2 < 0,$$

where the last inequality uses $\|\theta-\tilde{\theta}\| = \|\theta-\theta_0\| - \varepsilon \geq \varepsilon$. Hence,

$$\mathbb{P}\left(\forall\theta \text{ with } \|\theta-\theta_0\| \geq 2\varepsilon, G_N^\lambda(\theta) \neq 0\right) \to 1.$$

By an argument similar to Case 1, we conclude that $\hat{\theta}^\lambda \xrightarrow{p} \theta_0$.

A.6.3 Proof of Theorem 2

Building on the consistency established in Theorem 1, we derive the asymptotic distribution of $\hat{\theta}^\lambda$ following the approach of Theorem 5.21 in Van der Vaart (2000). The following proof is carried out on the event that $\hat{\theta}^\lambda$ is the unique solution to $G_N^\lambda(\theta) = 0$, which occurs with probability tending to one.

Since $G_N^\lambda(\hat{\theta}^\lambda) = G(\theta_0) = 0$, we can write

$$\begin{aligned} G(\theta_0) - G(\hat{\theta}^\lambda) = \; & \frac{1}{N_0}\sum_{i\in S_0}\left(g_{\hat{\theta}^\lambda}(X_i,Y_i) - \mathbb{E}[g_{\hat{\theta}^\lambda}(X_i,Y_i)]\right) \\ & + \sum_{k=1}^{K}\sum_{j=1}^{m}\lambda_j\frac{N_j}{N}\left[\frac{1}{N_j}\sum_{i\in S_{j,k}}\left(g_{\hat{\theta}^\lambda}(f_{-k}^j(X_i),Y_i) - \mathbb{E}_Z[g_{\hat{\theta}^\lambda}(f_{-k}^j(X_i),Y_i)]\right)\right. \\ & \left. - \frac{1}{N_0}\sum_{i\in S_{0,k}}\left(g_{\hat{\theta}^\lambda}(f_{-k}^j(X_i),Y_i) - \mathbb{E}_Z[g_{\hat{\theta}^\lambda}(f_{-k}^j(X_i),Y_i)]\right)\right], \end{aligned}$$

where $\mathbb{E}_Z[\cdot]$ denotes expectation over the data conditional on the imputation function.

By Assumption 2.2, $G(\theta)$ is differentiable at θ_0 with Jacobian J_{θ_0}, yielding the expansion

$$G(\theta_0) - G(\hat{\theta}^\lambda) = -J_{\theta_0}(\hat{\theta}^\lambda - \theta_0) + o_p(\|\hat{\theta}^\lambda - \theta_0\|).$$

The consistency of $\hat{\theta}^\lambda$ and the Lipschitz continuity of g_θ allows us to apply Lemma 19.24 from Van der Vaart (2000), which establishes that the empirical average evaluated at $\hat{\theta}^\lambda$ and θ_0 are asymptotically equivalent. Specifically, for the fully observed units,

$$\frac{1}{\sqrt{N_0}} \sum_{i \in S_{0,k}} \left(g_{\hat{\theta}^\lambda}(X_i, Y_i) - \mathbb{E}[g_{\hat{\theta}^\lambda}(X_i, Y_i)] - (g_{\theta_0}(X_i, Y_i) - \mathbb{E}[g_{\theta_0}(X_i, Y_i)]) \right) = o_p(1),$$

and similarly for partially observed units with $N_j \to \infty$,

$$\begin{aligned} \frac{1}{\sqrt{N_j}} \sum_{i \in S_{j,k}} & \left(g_{\hat{\theta}^\lambda}(f^j_{-k}(X_i), Y_i) - \mathbb{E}_Z[g_{\hat{\theta}^\lambda}(f^j_{-k}(X_i), Y_i)] \right. \\ & \left. - \left(g_{\theta_0}(f^j_{-k}(X_i), Y_i) - \mathbb{E}_Z[g_{\theta_0}(f^j_{-k}(X_i), Y_i)] \right) \right) = o_p(1). \end{aligned}$$

Furthermore, the stability condition (6) ensures that for each j with $N_j \to \infty$, there exists a deterministic function $\bar{f}^j$ such that

$$\begin{aligned} \frac{1}{\sqrt{N_j}} \sum_{i \in S_{j,k}} & \left(g_{\theta_0}(f^j_{-k}(X_i), Y_i) - \mathbb{E}_Z[g_{\theta_0}(f^j_{-k}(X_i), Y_i)] \right. \\ & \left. - \left(g_{\theta_0}(\bar{f}^j(X_i), Y_i) - \mathbb{E}[g_{\theta_0}(\bar{f}^j(X_i), Y_i)] \right) \right) = o_p(1). \end{aligned}$$

Combining these results and noting that groups with $N_j = O(1)$ contribute negligibly to the asymptotic distribution, we obtain

$$\begin{aligned} -J_{\theta_0}(\hat{\theta}^\lambda - \theta_0) = \; & o_p(\|\hat{\theta}^\lambda - \theta_0\|) + \frac{1}{N_0} \sum_{i \in S_0} (g_{\theta_0}(X_i, Y_i) - \mathbb{E}[g_{\theta_0}(X_i, Y_i)]) \\ & + \sum_{k=1}^{K} \sum_{j=1}^{m} \lambda_j \frac{N_j}{N} \left[\frac{1}{N_j} \sum_{i \in S_{j,k}} \left(g_{\theta_0}(\bar{f}^j(X_i), Y_i) - \mathbb{E}_Z[g_{\theta_0}(\bar{f}^j(X_i), Y_i)] \right) \right. \\ & \left. - \frac{1}{N_0} \sum_{i \in S_{0,k}} \left(g_{\theta_0}(\bar{f}^j(X_i), Y_i) - \mathbb{E}_Z[g_{\theta_0}(\bar{f}^j(X_i), Y_i)] \right) \right] + o_p\left(\frac{1}{\sqrt{N_0}} \right). \end{aligned}$$

The right-hand side consists of centered averages of i.i.d. random variables across independent

folds. Moreover, since J_{θ_0} is invertible, it holds that

$$\sqrt{N_0}\|\hat{\theta}^\lambda - \theta_0\| \leq \|J_{\theta_0}^{-1}\|\sqrt{N_0}\|J_{\theta_0}(\hat{\theta}^\lambda - \theta_0)\| = O_p(1) + o_p(\sqrt{N_0}\|\hat{\theta}^\lambda - \theta_0\|),$$

which implies $\|\hat{\theta}^\lambda - \theta_0\| = O_p(1/\sqrt{N_0})$.

Multiplying both sides by $\sqrt{N_0}J_{\theta_0}^{-1}$ and applying the central limit theorem yields

$$\sqrt{N_0}(\hat{\theta}^\lambda - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, J_{\theta_0}^{-1}\left(\Sigma_{\Delta,0}^\lambda + \sum_{j=1}^m \lambda_j^2 r_j r_0 \Sigma_{f,j}\right) J_{\theta_0}^{-1}\right),$$

with $\Sigma_{f,j}$ and $\Sigma_{\Delta,0}^\lambda$ as defined in the theorem. This completes the proof.

A.6.4 Proof of Corollary 1

Based on Theorem 2, the average asymptotic variance, that we want to minimize, is equal to

$$\mathrm{tr}(\Sigma^\lambda) = \mathrm{tr}\left(J_{\theta_0}^{-1}\left(\Sigma_{\Delta,0}^\lambda + \sum_{j=1}^m \lambda_j^2 r_j r_0 \Sigma_{f,j}\right) J_{\theta_0}^{-1}\right).$$

Expanding $\Sigma_{\Delta,0}^\lambda = \mathrm{Cov}(g_{\theta_0}(X_i, Y_i) - \sum_{j=1}^m \lambda_j r_j g_{\theta_0}^{(j)})$, we obtain

$$\begin{aligned}\Sigma_{\Delta,0}^\lambda = \ & \mathrm{Cov}(g_{\theta_0}) - \sum_{j=1}^m \lambda_j r_j [\mathrm{Cov}(g_{\theta_0}, g_{\theta_0}^{(j)}) + \mathrm{Cov}(g_{\theta_0}^{(j)}, g_{\theta_0})] \\ & + \sum_{j,h=1}^m \lambda_j \lambda_h r_j r_h \mathrm{Cov}(g_{\theta_0}^{(j)}, g_{\theta_0}^{(h)}).\end{aligned}$$

Plugging this and $\Sigma_{f,j} = \mathrm{Cov}(g_{\theta_0}^{(j)})$ into $\mathrm{tr}(\Sigma^\lambda)$, we have

$$\begin{aligned}\mathrm{tr}(\Sigma^\lambda) = \ & \mathrm{tr}\left(J_{\theta_0}^{-1}\mathrm{Cov}(g_{\theta_0})J_{\theta_0}^{-1}\right) - \sum_{j=1}^m \lambda_j r_j \mathrm{tr}\left(J_{\theta_0}^{-1}(\mathrm{Cov}(g_{\theta_0}, g_{\theta_0}^{(j)}) + \mathrm{Cov}(g_{\theta_0}^{(j)}, g_{\theta_0}))J_{\theta_0}^{-1}\right) \\ & + \sum_{j,h=1}^m \lambda_j \lambda_h r_j r_h \mathrm{tr}\left(J_{\theta_0}^{-1}\mathrm{Cov}(g_{\theta_0}^{(j)}, g_{\theta_0}^{(h)})J_{\theta_0}^{-1}\right) + \sum_{j=1}^m \lambda_j^2 r_j r_0 \mathrm{tr}\left(J_{\theta_0}^{-1}\mathrm{Cov}(g_{\theta_0}^{(j)})J_{\theta_0}^{-1}\right).\end{aligned}$$

Taking the derivative with respect to λ_j and setting it equal to zero, we get that $Q\lambda^* = v$. Since the objective function is quadratic in λ and Q is positive semi-definite, λ^* is indeed a global

minimum. To derive the optimal variance, we substitute λ^* into the objective:

$$\begin{aligned}\operatorname{tr}(\Sigma^{\lambda^*}) &= \operatorname{tr}\left(J_{\theta_0}^{-1}\operatorname{Cov}(g_{\theta_0})J_{\theta_0}^{-1}\right) - 2v^\top\lambda^* + \lambda^{*\top}Q\lambda^* \\ &= \operatorname{tr}\left(J_{\theta_0}^{-1}\operatorname{Cov}(g_{\theta_0})J_{\theta_0}^{-1}\right) - \lambda^{*\top}Q\lambda^* \leq \operatorname{tr}\left(J_{\theta_0}^{-1}\operatorname{Cov}(g_{\theta_0})J_{\theta_0}^{-1}\right),\end{aligned}$$

where the last inequality holds since Q is positive semi-definite.

A.6.5 Proof of Corollary 2

We have the decomposition

$$G_N^{\hat{\lambda}}(\theta_0) - G_N^{\lambda^*}(\theta_0) = \sum_{j=1}^{m}(\hat{\lambda}_j - \lambda_j^*)\, D_{j,N}(\theta_0),$$

where

$$D_{j,N}(\theta) = \sum_{k=1}^{K}\frac{N_j}{N}\left(\frac{1}{N_j}\sum_{i\in S_{j,k}} g_\theta(f_{-k}^j(X_i), Y_i) - \frac{1}{N_0}\sum_{i\in S_{0,k}} g_\theta(f_{-k}^j(X_i), Y_i)\right).$$

Under the conditions of Theorem 2, each $D_{j,N}(\theta_0) = O_p(N_0^{-1/2})$, hence

$$\sqrt{N_0}\left\|G_N^{\hat{\lambda}}(\theta_0) - G_N^{\lambda^*}(\theta_0)\right\| \leq \sum_{j=1}^{m}|\hat{\lambda}_j - \lambda_j^*| \cdot \sqrt{N_0}\,\|D_{j,N}(\theta_0)\| = o_p(1).$$

Following the proof of Theorem 1, we establish that $\sup_{\lambda\in\Lambda}\|\hat{\theta}^\lambda - \theta_0\| = o_p(1)$ for compact Λ, and thus following Theorem 2 we establish the linear expansion

$$-J_{\theta_0}(\hat{\theta}^\lambda - \theta_0) = G_N^\lambda(\theta_0) + o_p(N_0^{-1/2}),$$

uniformly over $\lambda \in \Lambda$. Applying this expansion at $\lambda = \hat{\lambda}$ and $\lambda = \lambda^*$ and multiplying by $\sqrt{N_0}$ yields

$$\sqrt{N_0}\,(\hat{\theta}^{\hat{\lambda}} - \theta_0) = -J_{\theta_0}^{-1}\sqrt{N_0}\,G_N^{\lambda^*}(\theta_0) + o_p(1),$$

which implies that $\sqrt{N_0}\,(\hat{\theta}^{\hat{\lambda}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma^{\lambda^*})$.

A.6.6 Proof of Proposition 2

Proof. Following Bayle, Bayle, Janson, and Mackey (2020), a sufficient condition for (6) is

$$\mathbb{E}_Z\left[\left\|g_{\theta_0}(f^j_{-k}(X_i), Y_i) - g_{\theta_0}(\bar{f}^j(X_i), Y_i)\right\|^2\right] \xrightarrow{p} 0.$$

Applying the mean continuity condition (9) with $X_1 = f^j_{-k}(X_i)$ and $X_2 = \bar{f}^j(X_i)$ yields

$$\mathbb{E}_Z\left[\left\|g_{\theta_0}(f^j_{-k}(X_i), Y_i) - g_{\theta_0}(\bar{f}^j(X_i), Y_i)\right\|^2\right] \lesssim \left(\mathbb{E}_X\left[\|f^j_{-k}(X_i) - \bar{f}^j(X_i)\|^q\right]\right)^{2/q}.$$

Therefore, it suffices to show that

$$\mathbb{E}_X\left[\|f^j_{-k}(X_i) - \bar{f}^j(X_i)\|^q\right] \xrightarrow{p} 0.$$

Let $\bar{f}^j(x) = \mathbb{E}[f^j_{-k}(x)]$, where the expectation is taken over the randomness in the training data used to construct f^j_{-k}. Using the decomposition (8),

$$f^j_{-k}(x) - \bar{f}^j(x) = \Delta^j_{-k}(x) - \mathbb{E}[\Delta^j_{-k}(x)].$$

Assumption (10) implies $\mathbb{E}[\|\Delta^j_{-k}(X_1)\|^q] = o(1)$, which further yields

$$\mathbb{E}_X\left[\|\Delta^j_{-k}(X_1)\|^q\right] = o_p(1),$$

since by Markov's inequality, for any $\varepsilon > 0$,

$$\mathbb{P}\Big(\mathbb{E}_X[\|\Delta^j_{-k}(X_1)\|^q] > \varepsilon\Big) \le \frac{\mathbb{E}[\|\Delta^j_{-k}(X_1)\|^q]}{\varepsilon} = o(1).$$

Therefore, by Minkowski's inequality and Jensen's inequality:

$$\begin{aligned}
\Big(\mathbb{E}_X[\|f^j_{-k}(X_i) - \bar{f}^j(X_i)\|^q]\Big)^{1/q} &= \Big(\mathbb{E}_X[\|\Delta^j_{-k}(X_i) - \mathbb{E}[\Delta^j_{-k}(X_i)]\|^q]\Big)^{1/q} \\
&\le \Big(\mathbb{E}_X[\|\Delta^j_{-k}(X_i)\|^q]\Big)^{1/q} + \Big(\mathbb{E}_X[\|\mathbb{E}[\Delta^j_{-k}(X_i)]\|^q]\Big)^{1/q} \\
&\le \Big(\mathbb{E}_X[\|\Delta^j_{-k}(X_i)\|^q]\Big)^{1/q} + \Big(\mathbb{E}[\|\Delta^j_{-k}(X_i)\|^q]\Big)^{1/q} \\
&= o_p(1).
\end{aligned}$$

□

A.6.7 Proof of Proposition 3

Let $\hat{\Sigma}_{-k}$ denote the covariance estimator constructed from the training sample $[N]\backslash\tilde{S}_k$, and let $\hat{D}_{-k}$ be the largest eigenvalue of $\hat{\Sigma}_{-k}/L$. Let $\hat{F}_{-k}$ denote the unit-norm leading eigenvector of $\hat{\Sigma}_{-k}$, with its sign chosen so that $\hat{F}_{-k}^\top F \geq 0$.

Lemma A.3. *Under the assumptions of Proposition 3, for any $p \geq 1$,*

$$\mathbb{E}[|\hat{D}_{-k} - D|^p] = o(1),$$

where $D = \sigma_\Lambda^2 + \sigma_\epsilon^2/L$.

Proof. To simplify the notation, we suppress the index $-k$ in this proof. Define the sphere $\Gamma = \{\lambda \in \mathbb{R}^{L\times 1} : \lambda^\top\lambda = L\}$ and

$$\begin{aligned} R(\lambda) &= \frac{1}{L^2}\lambda^\top \left((X \odot W)^\top (X \odot W) \odot \left[\frac{1}{|Q_{hl}|}\right]\right)\lambda, \\ R^*(\lambda) &= \frac{1}{L^2}\lambda^\top \left(\sigma_\Lambda^2 FF^\top + \sigma_\epsilon^2 I_L\right)\lambda. \end{aligned}$$

Then, $\hat{D} = \sup_{\lambda\in\Gamma} R(\lambda)$ and $D = \sup_{\lambda\in\Gamma} R^*(\lambda)$. For any $\lambda \in \Gamma$, the difference $R(\lambda) - R^*(\lambda)$ can be decomposed as

$$\begin{aligned} R(\lambda) - R^*(\lambda) = \; & \frac{1}{L^2}\sum_{h,l=1}^{L} \lambda_h\lambda_l F_h F_l \left(\frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}} \Lambda_i^2 - \sigma_\Lambda^2\right) \\ & + \frac{1}{L^2}\sum_{h,l=1}^{L} \lambda_h\lambda_l \frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}} (\Lambda_i F_l \epsilon_{ih} + \Lambda_i F_h \epsilon_{il}) \\ & + \frac{1}{L^2}\sum_{h,l=1}^{L} \lambda_l\lambda_h \left(\frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}} \epsilon_{ih}\epsilon_{il} - \sigma_\epsilon^2 I(h=l)\right). \end{aligned}$$

By Hölder's inequality, we have

$$\begin{aligned}\sup_{\lambda\in\Gamma}|R(\lambda)-R^*(\lambda)|\leq\ & \underbrace{\sup_{\lambda\in\Gamma}\left(\frac{1}{L^2}\sum_{h,l=1}^{L}\lambda_h^2\lambda_l^2\right)^{1/2}}_{=1}\left(\frac{1}{L^2}\sum_{h,l=1}^{L}F_h^2F_l^2\left(\frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}\Lambda_i^2-\sigma_\Lambda^2\right)^2\right)^{1/2}\\ &+\sup_{\lambda\in\Gamma}\left(\frac{1}{L^2}\sum_{h,l=1}^{L}\lambda_h^2\lambda_l^2\right)^{1/2}\left(\frac{1}{L^2}\sum_{h,l=1}^{L}\left(\frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}(\Lambda_iF_l\epsilon_{ih}+\epsilon_{il}\Lambda_iF_h)\right)^2\right)^{1/2}\\ &+\sup_{\lambda\in\Gamma}\left(\frac{1}{L^2}\sum_{h,l=1}^{L}\lambda_l^2\lambda_h^2\right)^{1/2}\left(\frac{1}{L^2}\sum_{h,l=1}^{L}\left(\frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}\epsilon_{ih}\epsilon_{il}-\sigma_\epsilon^2\cdot I(h=l)\right)^2\right)^{1/2}.\end{aligned}$$

Therefore,

$$\begin{aligned}\sup_{\lambda\in\Gamma}|R(\lambda)-R^*(\lambda)|^p\lesssim\ & \max_{h,l}\left|\frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}\Lambda_i^2-\sigma_\Lambda^2\right|^p+\max_{h,l}\left|\frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}\epsilon_{ih}\epsilon_{il}-\sigma_\epsilon^2\cdot I(h=l)\right|^p\\ &+\max_{h,l}\left|\frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}(\Lambda_iF_l\epsilon_{ih}+\epsilon_{il}\Lambda_iF_h)\right|^p.\end{aligned}$$

Since $\Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_\Lambda^2)$, $\epsilon_{il} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_\epsilon^2)$, $W \perp\!\!\!\perp \Lambda$, and $|Q_{hl}|/N \geq c > 0$, we have

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}\Lambda_i^2-\sigma_\Lambda^2\right|^p\right]=o(1),\\ &\mathbb{E}\left[\left|\frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}\epsilon_{ih}\epsilon_{il}-\sigma_\epsilon^2\cdot I(h=l)\right|^p\right]=o(1),\\ &\mathbb{E}\left[\left|\frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}(\Lambda_iF_l\epsilon_{ih}+\epsilon_{il}\Lambda_iF_h)\right|^p\right]=o(1).\end{aligned}$$

Furthermore, there are finitely many pairs (h,l). As a result, $\mathbb{E}\left[\sup_{\lambda\in\Gamma}|R(\lambda)-R^*(\lambda)|^p\right]=o(1)$, which implies that

$$\begin{aligned}\mathbb{E}\left[|\hat{D}-D|^p\right]=\ & \mathbb{E}\left[|\sup_{\lambda\in\Gamma}R(\lambda)-\sup_{\lambda\in\Gamma}R^*(\lambda)|^p\right]\\ \leq\ & \mathbb{E}\left[\sup_{\lambda\in\Gamma}|R(\lambda)-R^*(\lambda)|^p\right]=o(1).\end{aligned}$$

□

Lemma A.4. *Under the assumptions of Proposition 3, for any $p \geq 1$ and $l = 1, \ldots, L$,*

$$\mathbb{E}[|\hat{F}_{-k,l} - F_l|^p] = o(1).$$

Proof. We suppress the index $-k$ for the sake of notational simplicity. Since we estimate the factors by applying PCA to $\hat{\Sigma}$:

$$\frac{1}{L}\hat{\Sigma}\hat{F} = \hat{F}\hat{D},$$

the l-th component of the estimated factor $\hat{F}$ can be expanded as

$$\begin{aligned}
\hat{F}_l = \; & \hat{D}^{-1}\frac{1}{L}\sum_{h=1}^{L}\hat{F}_h F_h \frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}\Lambda_i^2 F_l + \hat{D}^{-1}\frac{1}{L}\sum_{h=1}^{L}\hat{F}_h F_l \frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}\Lambda_i\epsilon_{ih} \\
& + \hat{D}^{-1}\frac{1}{L}\sum_{h=1}^{L}\hat{F}_h F_h \frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}\Lambda_i\epsilon_{il} + \hat{D}^{-1}\frac{1}{L}\sum_{h=1}^{L}\tilde{F}_h \frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}\epsilon_{ih}\epsilon_{il}.
\end{aligned}$$

We define $H_l = \hat{D}^{-1}\frac{1}{L}\sum_{h=1}^{L}\hat{F}_h F_h \frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}\Lambda_i^2$ and $H = (NL)^{-1}\hat{D}^{-1}\hat{F}^\top F\Lambda^\top\Lambda$. Furthermore, to simplify the notation,

$$\begin{aligned}
\eta_{hl} = \; & \frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}} F_l\Lambda_i\epsilon_{ih}, \\
\xi_{hl} = \; & \frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}} F_h\Lambda_i\epsilon_{il}, \\
\gamma_{hl} = \; & \frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}} \epsilon_{ih}\epsilon_{il}.
\end{aligned}$$

Plugging these into $\hat{F}_l$, we have

$$\hat{F}_l = H_l F_l + \hat{D}^{-1}\frac{1}{L}\sum_{h=1}^{L}\hat{F}_h(\eta_{hl} + \xi_{hl} + \gamma_{hl}),$$

which implies that

$$\begin{aligned}
\hat{F}_l - \frac{1}{L}D^{-1}\hat{F}_l\sigma_\epsilon^2 - HF_l = \; & (H_l - H)F_l + \hat{D}^{-1}\frac{1}{L}\sum_{h=1}^{L}\hat{F}_h(\eta_{hl} + \xi_{hl} + \gamma_{hl}) - \frac{1}{L}D^{-1}\hat{F}_l\sigma_\epsilon^2 \\
= \; & (H_l - H)F_l + \hat{D}^{-1}\frac{1}{L}\sum_{h=1}^{L}\hat{F}_h\left(\eta_{hl} + \xi_{hl} + (\gamma_{hl} - \sigma_\epsilon^2 I(h=l))\right) \\
& + \frac{1}{L}(\hat{D}^{-1} - D^{-1})\hat{F}_l\sigma_\epsilon^2.
\end{aligned}$$

In the following, we bound the moment of each term on the right-hand side. First, it holds that

$$\hat{D} \geq \frac{1}{L}\sum_{i=1}^{L}\lambda_i\left(\frac{1}{L}\hat{\Sigma}\right) = \frac{1}{L}\mathrm{trace}\left(\frac{1}{L}\hat{\Sigma}\right) = \frac{1}{L^2}\sum_{l=1}^{L}\frac{1}{|Q_{ll}|}\sum_{i\in Q_{ll}}(\Lambda_i F_l + \epsilon_{il})^2,$$

where $\lambda_i(A)$ denotes the i-th largest eigenvalue of the matrix A. Observe that the terms on the right-hand side are chi-squared distributed, there exists a constant C such that $\mathbb{E}[|\hat{D}|^{-p}] \leq C$. By Lemma A.3, $\mathbb{E}[|\hat{D} - D|^q] = o(1)$ for any q, thus

$$\mathbb{E}\left[\left|\hat{D}^{-1} - D^{-1}\right|^p\right] \leq \left(\mathbb{E}\left[\frac{1}{(D\hat{D})^{2p}}\right]\cdot\mathbb{E}\left[(D-\hat{D})^{2p}\right]\right)^{1/2} = o(1).$$

Under the assumptions of the model, $\mathbb{E}\left[|\eta_{hl}|^p\right] = o(1)$, $\mathbb{E}\left[|\xi_{hl}|^p\right] = o(1)$, and $\mathbb{E}\left[|\gamma_{hl} - \sigma_\epsilon^2 I(h=l)|^p\right] = o(1)$ for any p. By the boundedness of $\hat{F}_h$, it holds that

$$\mathbb{E}\left[\left|\frac{1}{L}\sum_{h=1}^{L}\hat{F}_h(\eta_{hl} + \xi_{hl} + \gamma_{hl} - \sigma_\epsilon^2 I(h=l))\right|^p\right] = o(1).$$

Furthermore, note that

$$\begin{aligned}
\mathbb{E}\left[|H_l - H|^p\right] &= \mathbb{E}\left[\left|\frac{1}{L}\hat{D}^{-1}\sum_{h=1}^{L}\hat{F}_h F_h\left(\frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}\Lambda_i^2 - \frac{1}{N}\sum_{i=1}^{N}\Lambda_i^2\right)\right|^p\right] \\
&\lesssim \left(\mathbb{E}[\hat{D}^{-2p}]\cdot\frac{1}{L}\sum_{h=1}^{L}\mathbb{E}\left[\left(\frac{1}{|Q_{hl}|}\sum_{i\in Q_{hl}}\Lambda_i^2 - \frac{1}{N}\sum_{i=1}^{N}\Lambda_i^2\right)^{2p}\right]\right)^{1/2} \\
&= o(1),
\end{aligned}$$

where the last equality holds since $\Lambda_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\Lambda^2)$ and $W \perp\!\!\!\perp \Lambda$.

Combining these results, we get

$$\mathbb{E}[|\hat{F}_l - (LD)^{-1}\hat{F}_l\sigma_\epsilon^2 - HF_l|^p] = o(1).$$

Since $D = \sigma_\Lambda^2 + \sigma_\epsilon^2/L$, we have $1 - \sigma_\epsilon^2/(LD) = \sigma_\Lambda^2/D$, so that

$$\mathbb{E}\left[\left|\hat{F}_l - \frac{D}{\sigma_\Lambda^2}HF_l\right|^p\right] = o(1),$$

Let $G = DH/\sigma_\Lambda^2$. Observe that

$$\begin{aligned} o(1) = \mathbb{E}\left[\left(\frac{1}{L}\sum_{l=1}^{L}(\hat{F}_l - GF_l)^2\right)^p\right] &= \mathbb{E}\left[\left(\frac{1}{L}\sum_{l=1}^{L}(\hat{F}_l^2 - 2GF_l\hat{F}_l + G^2F_l^2)\right)^p\right] \\ &= \mathbb{E}\left[\left(1 - G^2 - \frac{2}{L}\sum_{l=1}^{L}GF_l(\hat{F}_l - GF_l)\right)^p\right] \\ &\geq C_1 \cdot \mathbb{E}\left[\left|1 - G^2\right|^p\right] - C_2 \cdot \underbrace{\mathbb{E}\left[\left|\frac{2}{L}\sum_{l=1}^{L}GF_l(\hat{F}_l - GF_l)\right|^p\right]}_{o(1)} \end{aligned}$$

for some positive constants C_1 and C_2. This implies $\mathbb{E}\left[\left|1 - G^2\right|^p\right] = o(1)$. Furthermore, since the signs of $\hat{F}$ align with those of F, we have $G > 0$ and $\mathbb{E}\left[|1 - G|^p\right] = o(1)$. As a result,

$$\begin{aligned} \mathbb{E}\left[|\hat{F}_l - F_l|^p\right] &= \mathbb{E}\left[|\hat{F}_l - GF_l + (G-1)F_l|^p\right] \\ &\lesssim \mathbb{E}\left[|\hat{F}_l - GF_l|^p\right] + \mathbb{E}\left[|(G-1)F_l|^p\right] = o(1). \end{aligned}$$

□

Proof. Now we verify condition (10) for the factor-based imputation. For any pattern j, fold k, and coordinate l, the imputed value $f^j_{-k,l}(x)$ can be written as

$$\begin{aligned} f^j_{-k,l}(x) = &\left(\sum_{h=1}^{L} W_h^{(j)}\hat{F}^2_{-k,h}\right)^{-1} \cdot \sum_{h=1}^{L} W_h^{(j)}\hat{F}_{-k,h}\hat{F}_{-k,l}x_h \\ = &\left(\sum_{h=1}^{L} W_h^{(j)}F_h^2\right)^{-1}\sum_{h=1}^{L} W_h^{(j)}F_hF_lx_h + \left(\sum_{h=1}^{L} W_h^{(j)}F_h^2 - \sum_{h=1}^{L} W_h^{(j)}\hat{F}^2_{-k,h}\right)\cdot \\ &\left(\sum_{h=1}^{L} W_h^{(j)}F_h^2\sum_{h=1}^{L} W_h^{(j)}\hat{F}^2_{-k,h}\right)^{-1}\sum_{h=1}^{L} W_h^{(j)}F_hF_lx_h \\ &+ \left(\sum_{h=1}^{L} W_h^{(j)}\hat{F}^2_{-k,h}\right)^{-1}\sum_{l=1}^{L} W_h^{(j)}(\hat{F}_{-k,h}\hat{F}_{-k,l} - F_hF_l)x_h. \end{aligned}$$

We define $h_l^j(x) = \left(\sum_{h=1}^L W_h^{(j)} F_h^2\right)^{-1} \sum_{h=1}^L W_h^{(j)} F_h F_l x_h$ in decomposition (8), and then

$$\Delta_{-k,l}^j(x) = \underbrace{\left(\sum_{h=1}^L W_h^{(j)} F_h^2 - \sum_{h=1}^L W_h^{(j)} \hat{F}_{-k,h}^2\right)\left(\sum_{h=1}^L W_h^{(j)} F_h^2 \sum_{h=1}^L W_h^{(j)} \hat{F}_{-k,h}^2\right)^{-1} \sum_{h=1}^L W_h^{(j)} F_h F_l x_h}_{\Delta_{1,-k,l}^j(x)}$$
$$+ \underbrace{\left(\sum_{h=1}^L W_h^{(j)} \hat{F}_{-k,h}^2\right)^{-1} \sum_{h=1}^L W_h^{(j)} (\hat{F}_{-k,h}\hat{F}_{-k,l} - F_h F_l) x_h}_{\Delta_{2,-k,l}^j(x)}.$$

In the following, we bound the moments of $\Delta_{1,-k,l}^j(X_i)$ and $\Delta_{2,-k,l}^j(X_i)$, respectively.

Since the estimated factors $\hat{F}_{-k}$ are independent of X_i and $\sum_{h=1}^L W_h^{(j)} F_h^2 \geq \min_h F_h^2 > 0$,

$$\mathbb{E}[|\Delta_{1,-k,l}^j(X_i)|^q] \lesssim \mathbb{E}\left[\left|\sum_{h=1}^L W_h^{(j)}(F_h^2 - \hat{F}_{-k,h}^2)\right|^q \left|\sum_{h=1}^L W_h^{(j)} \hat{F}_{-k,h}^2\right|^{-q}\right] \cdot \mathbb{E}\left[\left|\sum_{h=1}^L W_h^{(j)} F_h F_l X_{ih}\right|^q\right]$$
$$\leq \left(\mathbb{E}\left[\left|\sum_{h=1}^L W_h^{(j)}(F_h^2 - \hat{F}_{-k,h}^2)\right|^{2q}\right]\right)^{1/2} \underbrace{\left(\mathbb{E}\left[\left|\sum_{h=1}^L W_h^{(j)} \hat{F}_{-k,h}^2\right|^{-2q}\right]\right)^{1/2}}_{O(1)} \cdot$$
$$\underbrace{\mathbb{E}\left[\left|\sum_{h=1}^L W_h^{(j)} F_h F_l X_{ih}\right|^q\right]}_{O(1)}.$$

By Lemma A.4, for any $p \geq 1$, we have $\mathbb{E}[|\hat{F}_{-k,l} - F_l|^p] = o(1)$, and thus

$$\mathbb{E}\left[\left|(F_h^2 - \hat{F}_{-k,h}^2)\right|^{2q}\right] \leq \left(\mathbb{E}\left[(F_h - \hat{F}_{-k,h})^{4q}\right]\right)^{1/2} \cdot \left(\mathbb{E}\left[(F_h + \hat{F}_{-k,h})^{4q}\right]\right)^{1/2} = o(1),$$

which implies that $\mathbb{E}[|\Delta_{1,-k,l}^j(X_i)|^q] = o(1)$. Similarly, we can prove $\mathbb{E}[|\Delta_{2,-k,l}^j(X_i)|^q] = o(1)$. Summing over l yields $\mathbb{E}[\|\Delta_{-k}^j(X_i)\|^q] = o(1)$, which establishes condition (10) and completes the proof. □

A.6.8 Proof of Theorem 4

Following the proof of Theorem 1, we first establish the consistency of the estimator $\hat{\theta}^\lambda$.

In the following, we establish the asymptotic normality of $\hat{\theta}^{\lambda}$. For any $l = 1, \ldots, L$, we can write

$$
\begin{aligned}
& G_l(\theta_0) - G_l(\hat{\theta}^{\lambda}) \\
= & \sum_{h=1}^{L} \frac{1}{|Q_{hl}|} \sum_{i \in Q_{hl}} \left(X_{il} X_{ih} \hat{\theta}_h^{\lambda} - \mathbb{E}[X_{il} X_{ih} \hat{\theta}_h^{\lambda}] \right) - \frac{1}{|Q_{ll}|} \sum_{i \in Q_{ll}} (X_{il} Y_i - \mathbb{E}[X_{il} Y_i]) \\
& + \sum_{k=1}^{K} \sum_{h=1}^{L} \sum_{j \notin I_{hl}} \lambda_j \frac{N_j}{N} \left(\frac{1}{N_j} \sum_{i \in S_{j,k}} (f_{-k,l}^{j}(X_i) f_{-k,h}^{j}(X_i) \hat{\theta}_h^{\lambda} - \mathbb{E}_Z[f_{-k,l}^{j}(X_i) f_{-k,h}^{j}(X_i) \hat{\theta}_h^{\lambda}]) \right. \\
& \left. - \frac{1}{|Q_{hl}|} \sum_{j' \in I_{hl}} \sum_{i \in S_{j',k}} (f_{-k,l}^{j}(X_i) f_{-k,h}^{j}(X_i) \hat{\theta}_h^{\lambda} - \mathbb{E}_Z[f_{-k,l}^{j}(X_i) f_{-k,h}^{j}(X_i) \hat{\theta}_h^{\lambda}]) \right) - \sum_{k=1}^{K} \sum_{j \notin I_{ll}} \lambda_j \frac{N_j}{N} \cdot \\
& \left(\frac{1}{N_j} \sum_{i \in S_{j,k}} (f_{-k,l}^{j}(X_i) Y_i - \mathbb{E}[f_{-k,l}^{j}(X_i) Y_i]) - \frac{1}{|Q_{ll}|} \sum_{j' \in I_{ll}} \sum_{i \in S_{j',k}} (f_{-k,l}^{j}(X_i) Y_i - \mathbb{E}[f_{-k,l}^{j}(X_i) Y_i]) \right),
\end{aligned}
$$

where the imputation functions are given by $f_{-k}^{j}(\cdot) = f^{j}(\cdot; [N] \backslash \tilde{S}_k)$.

We define

$$
\begin{aligned}
\tilde{\varphi}_{j,i,l} = & \sqrt{N_* N_j} \left[\sum_{h=1}^{L} I(j \in I_{hl}) \frac{1}{|Q_{hl}|} \left(X_{il} X_{ih} \theta_{0,h} - \sum_{j' \notin I_{hl}} \lambda_{j'} \frac{N_{j'}}{N} \bar{f}_l^{j'}(X_i) \bar{f}_h^{j'}(X_i) \theta_{0,h} \right) \right. \\
& \left. + I(j \notin I_{hl}) \lambda_j \frac{1}{N} \bar{f}_l^{j}(X_i) \bar{f}_h^{j}(X_i) \theta_{0,h} \right] - \sqrt{N_* N_j} \left[I(j \in I_{ll}) \frac{1}{|Q_{ll}|} \left(X_{il} Y_i - \sum_{j' \notin I_{ll}} \lambda_{j'} \frac{N_{j'}}{N} \bar{f}_l^{j'}(X_i) Y_i \right) \right. \\
& \left. + I(j \notin I_{ll}) \lambda_j \frac{1}{N} \bar{f}_l^{j}(X_i) Y_i \right].
\end{aligned}
$$

Proceeding as in the proof of Theorem 2, we obtain

$$
\sqrt{N_*}(G_l(\theta_0) - G_l(\hat{\theta}^{\lambda})) = \sum_{j=1}^{m} \frac{1}{\sqrt{N_j}} \sum_{i \in S_j} (\tilde{\varphi}_{j,i,l} - \mathbb{E}[\tilde{\varphi}_{j,i,l}]) + o_p(1).
$$

Let $\tilde{\varphi}_{j,i} = [\tilde{\varphi}_{j,i,1}; \ldots; \tilde{\varphi}_{j,i,L}] \in \mathbb{R}^L$ denote the vector formed by stacking the component-wise terms. Moreover, we have

$$
-J(\hat{\theta}^{\lambda} - \theta_0) = o_p(\|\hat{\theta}^{\lambda} - \theta_0\|) + \frac{1}{\sqrt{N_*}} \sum_{j=1}^{m} \frac{1}{\sqrt{N_j}} \sum_{i \in S_j} (\tilde{\varphi}_{j,i} - \mathbb{E}[\tilde{\varphi}_{j,i}]) + o_p(1) + o_p\left(\frac{1}{\sqrt{N_*}} \right).
$$

Suppose that, as $N \to \infty$, $\sqrt{N_* N_j}/|Q_{hl}| \to r_{*,j,hl}$ and $\sqrt{N_* N_j}/N \to r_{*,j}$. Define

$$\varphi_{j,i,l} = \sum_{h=1}^{L} I(j \in I_{hl}) r_{*,j,hl} \left(X_{il} X_{ih} \theta_{0,h} - \sum_{j' \notin I_{hl}} \lambda_{j'} \frac{N_{j'}}{N} \bar{f}_l^{j'}(X_i) \bar{f}_h^{j'}(X_i) \theta_{0,h} \right) + I(j \notin I_{hl}) \lambda_j r_{*,j} \bar{f}_l^j(X_i) \cdot$$
$$\bar{f}_h^j(X_i) \theta_{0,h} - \left[I(j \in I_{ll}) r_{*,j,ll} \left(X_{il} Y_i - \sum_{j' \notin I_{ll}} \lambda_{j'} \frac{N_{j'}}{N} \bar{f}_l^{j'}(X_i) Y_i \right) + I(j \notin I_{ll}) \lambda_j r_{*,j} \bar{f}_l^j(X_i) Y_i \right].$$

Following the proof of Theorem 2, we conclude that

$$\sqrt{N_*}(\hat{\theta}^\lambda - \theta_0) \to \mathcal{N}\left(0, J^{-1} \sum_{j=1}^{m} \mathrm{Cov}(\varphi_{j,i}) J^{-1}\right).$$

A.6.9 Proof of Theorem 3

In this setting, the population moment is

$$G(\theta) = \sum_{j=0}^{m} \frac{N_j}{N} \mathbb{E}_{P^j}[g_\theta(X_i, Y_i)] = \sum_{j=0}^{m} \frac{N_j}{N} \mathbb{E}_{P^0}\left[g_\theta(X_i, Y_i) \omega^j(O_i)\right],$$

where $P^j = P_O^j \times P_{(X,Y)|O}$. The difference between the reweighted empirical moment $G_N^\lambda(\theta)$ from (11) and $G(\theta)$ can be decomposed as:

$$G_N^\lambda(\theta) - G(\theta) = \sum_{j=0}^{m} \frac{N_j}{N} \cdot \frac{1}{N_0} \sum_{i \in S_0} \left(g_\theta(X_i, Y_i) \omega^j(O_i) - \mathbb{E}_{P_0}[g_\theta(X_i, Y_i) \omega^j(O_i)] \right)$$
$$+ \sum_{k=1}^{K} \sum_{j=1}^{m} \lambda_j \frac{N_j}{N} \left[\frac{1}{N_j} \sum_{i \in S_{j,k}} \left(g_\theta(f_{-k}^j(X_i), Y_i) - \mathbb{E}_{Z \sim P^j}[g_\theta(f_{-k}^j(X_i), Y_i)] \right) \right.$$
$$\left. - \frac{1}{N_0} \sum_{i \in S_{0,k}} \left(g_\theta(f_{-k}^j(X_i), Y_i) \omega^j(O_i) - \mathbb{E}_{Z \sim P^0}[g_\theta(f_{-k}^j(X_i), Y_i) \omega^j(O_i)] \right) \right].$$

Following the proof of Theorem 1, we can first prove the consistency of the estimator $\hat{\theta}^\lambda$.

By Assumption 2.2, G is differentiable at θ_0 with Jacobian $J_{\theta_0} = \nabla G(\theta_0)$. We have

$$G(\theta_0) - G(\hat{\theta}^\lambda) = -J_{\theta_0}(\hat{\theta}^\lambda - \theta_0) + o_p(\|\hat{\theta}^\lambda - \theta_0\|).$$

Following similar arguments as in Theorem 2, it holds that

$$\begin{aligned}
-J_{\theta_0}(\hat{\theta}^\lambda - \theta_0) = \; & \sum_{j=0}^{m} \frac{N_j}{N} \cdot \frac{1}{N_0} \sum_{i \in S_0} \left(g_\theta(X_i, Y_i)\omega^j(O_i) - \mathbb{E}_{P_0}[g_\theta(X_i, Y_i)\omega^j(O_i)] \right) \\
& + \sum_{k=1}^{K} \sum_{j=1}^{m} \lambda_j \frac{N_j}{N} \left[\frac{1}{N_j} \sum_{i \in S_{j,k}} \left(g_{\theta_0}(\bar{f}^j(X_i), Y_i) - \mathbb{E}_{Z \sim P^j}[g_{\theta_0}(\bar{f}^j(X_i), Y_i)] \right) \right. \\
& \left. - \frac{1}{N_0} \sum_{i \in S_{0,k}} \left(g_{\theta_0}(\bar{f}^j(X_i), Y_i)\omega^j(O_i) - \mathbb{E}_{Z \sim P^0}[g_{\theta_0}(\bar{f}^j(X_i), Y_i)\omega^j(O_i)] \right) \right] \\
& + o_p(N_0^{-1/2}).
\end{aligned}$$

Let $r_j = \lim_{N \to \infty} N_j/N$ and define

$$\begin{aligned}
\Sigma_{\Delta,0}^{\lambda} = \; & \mathrm{Cov}_{P^0} \left(\sum_{j=0}^{m} r_j g_{\theta_0}(X_i, Y_i)\omega^j(O_i) - \sum_{j=1}^{m} \lambda_j r_j g_{\theta_0}(\bar{f}^j(X_i), Y_i)\omega^j(O_i) \right), \\
\Sigma_{f,j} = \; & \mathrm{Cov}_{P^j} \left(g_{\theta_0}(\bar{f}^j(X_i), Y_i) \right).
\end{aligned}$$

Applying the central limit theorem then yields

$$\sqrt{N_0}(\hat{\theta}^\lambda - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, J_{\theta_0}^{-1} \left(\Sigma_{\Delta,0}^{\lambda} + \sum_{j=1}^{m} \lambda_j^2 r_j r_0 \Sigma_{f,j} \right) J_{\theta_0}^{-1} \right).$$