Addressing Bias from Misclassification in Predicted Data: Theory and Application to Satellite Crop Data*

For most recent version, Click Here

Arthur R. Wardle†

October 2026

Abstract

Predicted data, generated by machine learning or other algorithms rather than direct measurement, are increasingly popular in empirical research but prone to measurement error. I introduce a flexible, nonparametric multiple imputation approach that uses a small validation sample with known true values to correct for misclassification in predicted categorical data, which would otherwise cause bias in model estimation. The method draws multiple plausible realizations of the unobserved true categories and estimates the model of interest on each completed dataset. I show that the simple average of these estimates is consistent for the true parameter of interest and apply standard multiple imputation confidence intervals for inference. Simulations confirm the procedure effectively removes bias under known data-generating processes. I then apply the procedure to a multinomial crop choice model for Kern County, California. Whereas naive use of satellite-based predicted data results in substantial bias, the multiple imputation method recovers estimates statistically indistinguishable from those based on full ground truth observations.

*I thank Maximilian Auffhammer, Ellen Bruno, Tamma Carleton, Joel Ferguson, Simon Greenhill, Nick Hagerty, James Sallee, Aaron Smith, and participants at the 2024 AAEA, 2025 TWEEDS and WEA, and 2026 IHS Junior Fellowship conferences for helpful feedback. I am grateful to Luna Yue Huang for helpful code. This research is supported by the Foundation for Food & Agriculture Research (FFAR) under Grant ID NIA21-0000000036 and from the Institute for Humane Studies under Grant ID IHS020193. Previous versions of this work circulated under the title "Model Estimation using Categorical Satellite Data with Misclassification."

†Department of Agricultural Economics & Economics, Montana State University, arthur.wardle@berkeley.edu

Introduction

Many new data sources for economically important variables now consist of *predicted data*: the output of statistical or machine learning models rather than direct measurements. Machine classification enables low-cost and scalable labeling of otherwise prohibitively large raw datasets. For example, machine classification of natural language is increasingly used for measuring features like topic or sentiment in speeches or online posts in political science and economics (de Slegte et al., 2024). A particularly widespread form of predicted data is *satellite data*, which extracts variables from satellite imagery using models trained on a small set of pre-labeled observations. Across agricultural, environmental, and development economics, satellite-based categorical data on forest cover, crop choice, burned areas, agricultural practice adoption, and urban land use are already playing a major role in applied research (Donaldson and Storeygard, 2016; Wuepper et al., 2025). As deep learning methods continue to improve, the availability and use of predicted data are likely to expand further (Dell, 2025).

Although these emerging data sources open new avenues for research, they require careful treatment because of potential measurement error. In particular, predicted data often involve categorical variables—such as land cover or crop type—that are subject to misclassification, meaning than an observation is assigned to the wrong category. Predicted data are typically more prone to misclassification than data obtained through traditional measurement. Such misclassification may be systematically related to other variables in a model, introducing a clear source of bias. However, even when misclassification is completely random, it remains an inherently non-classical form of measurement error that distorts parameter estimates.

Existing econometric methods for addressing measurement error are primarily designed for continuous variables, and those that are not generally require assumptions ill-suited to the predicted data context. These include parametric restrictions on the error process, valid instruments, or multiple proxies, sometimes in combination with validation data. The nature of misclassification in predicted data makes such assumptions difficult to justify. This paper

introduces a method that instead relies solely on validation data and minimal regularity assumptions.

The method I propose is a general purpose tool for end users of predicted data products, designed to remove bias from model estimates without reinventing the classification algorithm or requiring modification of standard statistical software. The approach starts by using a relatively small validation sample to learn the structure of misclassification in the predicted data—for example, that corn is more likely to be misclassified as wheat than apples, or that misclassification rates differ between irrigated and non-irrigated fields. This step is implemented nonparametrically to avoid misspecification bias.

Next, adapting techniques from the literature on *multiple imputation* for missing data, I generate many "corrected" datasets that represent plausible realizations of the unobserved true data, given what I learned about the structure of misclassification in the first step. The model of interest is then estimated separately on each "corrected" dataset, and the resulting estimates are combined using standard multiple imputation rules. The final estimator, the simple average of the estimates from each "corrected" dataset, is consistent for the parameter that would have been obtained using correctly classified data, with valid confidence intervals for inference.

More concretely, suppose the goal is to estimate a parameter β, using an estimator $\hat{\beta}(X, Z)$ that would satisfy $\mathbb{E}(\hat{\beta}(X, Z)) \approx \beta$ if the true covariates X and Z were observed. Instead, Z is unobserved for most units, and only a predicted proxy $\tilde{Z}$ is available. A small validation subsample contains both $\tilde{Z}$ and the true Z. Using the the law of iterated expectations, $\mathbb{E}(\hat{\beta}(X, Z)) = \mathbb{E}\left(\mathbb{E}\left(\hat{\beta}(X, Z)|X, \tilde{Z}\right)\right)$, multiple imputation estimates $\hat{\beta}(X, Z)$ by averaging over the estimates obtained from many simulated versions of Z. Each simulated Z is drawn from an *imputation model*, a model for $P(Z \mid \tilde{Z}, X)$ estimated using the validation data.

I evaluate the performance of the proposed method in a series of simulation exercises in which misclassification is introduced under a known data generating process. I then apply

the approach to estimating a multinomial crop choice model for Kern County, California, a major agricultural region at the southern end of California's Central Valley. Using both a comprehensive ground truth dataset and a satellite-derived predicted dataset for this area, I show that the method substantially mitigates the misclassification-induced biases observed in estimates that naively use the satellite data.

The remainder of the paper proceeds by discussing the sources of misclassification in predicted data, its consequences for estimation, and existing approaches for addressing it. I then introduce the general framework for multiple imputation and present a flexible, nonparametric implementation based on a hierarchical Dirichlet process mixture model. I then report results from simulation studies and apply the method to the Kern County crop choice data.

The Sources of Misclassification in Predicted Data

Predicted data, also referred to as *modeled* or *algorithmic* data, is data whose values are inferred by an algorithm rather than obtained through direct measurement. Continuing with the example of satellite crop data, a machine learning classifier (such as a random forest or decision tree) predicts which crop is grown in each pixel of a raster image or in each field of a shapefile dataset using satellite imagery. Such prediction algorithms vary widely in form and complexity, but they generally offer no guarantees that errors in the predicted data are independent of the true variable, other model variables, or of unobserved factors.

Figure 1 illustrates the provenance of predicted data and how it enters model estimation. Misclassification arises through several distinct channels.

First, errors in the raw input data propagate to predictions. For satellite data, these errors can come from imperfect sensors, distortions caused by cloud cover, smoke, or topography, errors in pre-processing that make imperfect adjustments for these issues, and errors introduced through interpolation to make up for incomplete satellite coverage.[1]

[1] Jain (2020) and Alix-Garcia and Millimet (2023) offer more thorough descriptions of how errors in the raw satellite data can impact downstream economics applications.

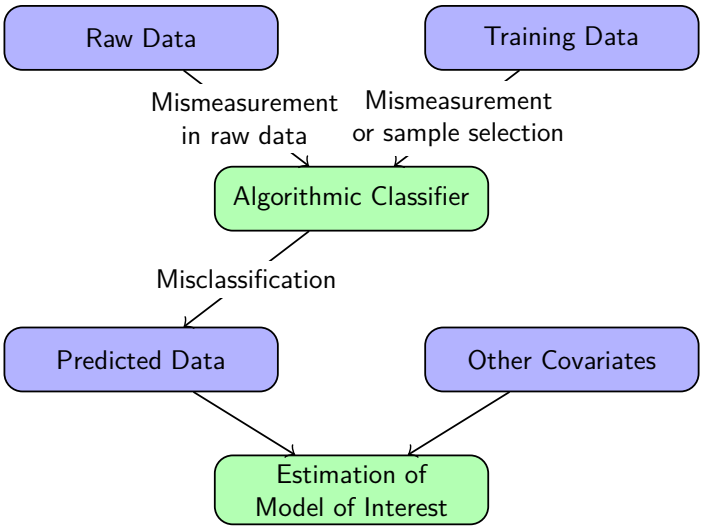


Figure 1: Origin and End Use of Predicted Data

Second, mismeasurement or sample selection in the training data can bias predictions. For example, the Cropland Data Layer, a satellite crop dataset used in the empirical application of this paper, relies on administrative data from federal farm support programs, oversampling crops that receive government support. Training on non-random samples introduces systematic bias into the predicted data.

Third, the algorithm itself produces prediction errors. Machine learning models are generally "black boxes," capable of achieving high degrees of predictive accuracy but without guarantees that remaining misclassification is independent of other variables.[2] Biases are often baked in during model development to seek greater predictive accuracy in the face of a bias-variance tradeoff (Dell, 2025). Only some predicted datasets offer any uncertaintly quantification, and those that do typically offer it in the form of a confusion matrix that summarizes misclassification rates across classes.[3] Short of perfect prediction accuracy, some degree

[2] Sanford et al. (2025) offer a technique for applying "adversarial debiasing" at the classifier training stage to ensure that errors in modeled outputs are uncorrelated with all other model covariates. This technique works when the satellite data are being used as a dependent variable and the classifier uses stochastic gradient descent for optimization. This requires the classification model to be tailored to a specific research design.

[3] For a classifier with J classes, a confusion matrix is a $J \times J$ matrix giving counts among all observations in some ground truth sample (sometimes the same one used for training, sometimes a holdout sample) with the true class along the x-axis and the algorithm-assigned class on the y-axis. A perfect classifier would produce a diagonal matrix.

of random error is inevitable; at best, this error is completely independent other model covariates.

Even widely used satellite crop datasets with reported overall accuracies in the neighborhood of 90 percent still contain enough misclassification to matter for econometric analysis. Despite this, predicted datasets are increasingly employed in applied economics and other fields without systematic assessment of misclassification or strategies to mitigate its impact on estimation.

Estimation Problems Under Misclassification

In linear regression with classical measurement error (that is, mean zero measurement error that is independent of the true variable and other covariates), the impact of measurement error depends on which variable is mismeasured. When the dependent variable is mismeasured, coefficient estimates remain unbiased, though variance increases. If an independent variable is mismeasured, its coefficient is attenuated and inconsistent, and coefficients on correlated variables will also be biased. These results, which are widely known and frequently form the basis for economists' intuitions about the impact of measurement error, do not apply to non-classical measurement error, nor do they apply to non-linear models.

Misclassification refers specifically to measurement error in categorical variables and is inherently non-classical. Unlike classical measurement error, misclassification in categorical data guarantees correlation between the measurement error and the true value, even if the error is completely random. This property means standard intuitions from classical measurement error do not apply to misclassification.

To see why, consider a binary variable $x \in \{0, 1\}$ and a randomly misclassified version $\tilde{x}$. The misclassification error $\eta = \tilde{x} - x$ can only take values in $\{-1, 0, 1\}$; in particular, when $x = 1$, $\eta \in \{-1, 0\}$ and when $x = 0$, $\eta \in \{0, 1\}$. This induces a negative correlation between x and η, no matter how misclassification arises. All mismeasurement in any discrete variable

exhibits this non-classical behavior.

Even small degress of misclassification can have large consequences. An early contribution by Aigner (1973) demonstrated that in simple linear regression with a single binary independent variable, completely random misclassification induces attenuation proportional to the misclassification rate. For non-linear models or dependent misclassification, it is difficult or impossible to derive formulas for how exactly misclassification will affect estimates. Simulation evidence confirms that even modest misclassification can produce extreme errors in parameter estimates (Millimet, 2010). As one example from the empirical literature, correcting for misclassification in employment status reversed a key result on fertility and female labor supply, despite an overall misclassification rate below 2 percent (Keane and Sauer, 2009).

In short, misclassification can substantially bias parameter estimates, particularly in non-linear models. This is true even when misclassification is completely random; potential dependence between misclassification in predicted data and other model variables only accentuates the problem. This motivates the need for a flexible correction method that can produce consistent estimates from predicted data subject to misclassification.

Existing Approaches

Identification under misclassification must rely on either additional structural assumptions or auxiliary data such as instruments, repeated measurements, or validation data.

One structural route to identification imposes restrictions on how misclassification arises. For instance, Hausman et al. (1998) show that for a categorical outcome model estimated via maximum likelihood, assuming that misclassification probabilities are independent of all other covariates ($\mathbb{P}(y = i|\tilde{y} = j, x) = \phi_{ij} \; \forall \; i, j$) enables consistent estimation through a simple adjustment to the likelihood function. However, this assumption is restrictive: even small violations can yield parameter estimates that differ significantly from those obtained using correctly classified data (Kreider, 2010).

Alix-Garcia and Millimet (2023) extend the Hausman et al. (1998) approach by allowing misclassification probabilities to depend on covariates, $\mathbb{P}(y = i|\tilde{y} = j, x, z) = G_{ij}(z'\beta)$, but this comes at the cost of requiring specific link functions G_{ij}. While this relaxes independence assumptions and is estimable without validation data, the true functional form of misclassification probabilities is generally unknown, even when researchers can identify which variables ought to be included in these functions.[4] It is not the case that *any* choice of function will alleviate bias either; Meyer and Mittag (2017) shows that misspecified misclassification functions can aggravate bias relative to a naive, uncorrected model. Because the algorithms underlying predicted data are often "black boxes" with unknown error-generating processes, strong functional form assumptions are generally implausible.[5]

An alternate route uses auxiliary data, such as instruments, repeated proxies, or validation samples. Methods using instruments often assume misclassification is independent of covariates (e.g. Frazis and Loewenstein, 2003), but Mahajan (2006) and Hu (2008) allow (in the binary and general discrete case respectively) for point identification even when misclassification probabilities are functions of observed data in the presence of an "instrument-like variable" that is correlated with the true underlying variable of interest but not the misclassification probability.[6] Predicted data will generally violate these requirements, since predicted data are a post-outcome noisy measurement of the true class, and therefore anything affecting the true class can also impact misclassification.[7] This literature continues to be active, offering

[4] See Mikhaeil et al. (forthcoming) for some specific cases where it is feasible to make reasonable assumptions about the functional forms of misclassification probabilities and how to estimate them in hierarchical Bayesian models.

[5] In cases where *all* variables in a model are categorical, Pepe and Fleming (1991) show how to estimate a nonparametric misclassification function to correct a maximum likelihood estimator making use of the empirical distribution of a validation sample, but this method quickly succumbs to the curse of dimensionality.

[6] Though DiTraglia and García-Jimeno (2019) show the Mahajan (2006) result to work only with exogeneity in the true value of the misclassified variable, with multiple "instrument-like" variables, or with multiple misclassified proxies.

[7] Rambachan et al. (2025) make a similar point to argue that surrogacy methods are inappropriate for satellite mismeasurement problems. However, these techniques can still be useful for statistics where correlation with other variables is not a concern. Torchiana et al. (2025) offer a specific implementation of the method from Hu (2008) to back out corrected transition probabilites using a Hidden Markov Model applied to a panel of satellite land classification data.

new approaches for point identification based on different moment restrictions, combinations of instruments, or combinations of misclassified proxies (Card, 1996; Lewbel, 2007; Battistin et al., 2014; DiTraglia and García-Jimeno, 2019; Nguimkeu et al., 2019; Yanagi, 2019; Calvi et al., 2022; Kasahara and Shimotsu, 2022). These approaches require either instruments, multiple misclassified proxies, or both for every observation.

Other work avoids the need for additional data by imposing restrictions on the relationship between misclassification and model error (Chen et al., 2008a,b, 2009). Chen et al. (2009) achieve nonparametric identification of a regression model with a misclassified covariate, relying on monotonicity of the regression function and conditional independence of both the true and misclassified covariate with the model error. Estimation, however, requires solving a complex system of equations, and the method does not apply to misclassified outcomes.

The method I propose relies only on a small, appropriately sampled validation dataset. It applies whether the misclassified variable is the outcome or an independent variable in the model of interest, and works for a wide class of estimators including OLS, MLE, GMM, and more. Moreover, the method does not require modification of an objective function, or anything else that would frustrate a practitioner's ability to use standard statistical estimation packages.

Proctor et al. (2023) also use multiple imputation, showing that it is highly effective for handling measurement error in continuous, remotely sensed data. They use a Bayesian linear regression for the imputation model, which does not accommodate imputation of categorical variables. My approach extends this idea to categorical data, allowing for nonparametric modeling of misclassification without relying on linearity in the imputation model.[8]

While others in the literature have sometimes considered the possibility of using multiple imputation to resolve misclassification in predicted data, they have generally written off the

[8]I will focus exclusively on the categorical case in this paper given its novelty, but the nonparametric imputation model does additionally allow for multiple imputation of continuous variables without the imposed linearity of the imputation model adopted by Proctor et al. (2023). In practice, imputation of continuous variables tends not to be overly sensitive to imputation model structure.

possibility due to potential bias from misspecification in the imputation model (e.g. Fong and Tyler, 2021; Kluger et al., 2025). Most implementations of multiple imputation in other contexts have historically relied on simple, parametric imputation models, since they were designed for problems where minor misspecification would have only minor consequences for downstream analysis. For misclassification in predicted data, where validation data may exist for only a small share of the overall dataset, relying on this logic is riskier, since misspecification of the imputation model will impact a large share of observations. For this reason, I apply recent advancements in nonparametric Bayesian estimation for estimating the imputation model, and offer a novel proof that this guarantees consistency of the multiple imputation estimator.

Two other emerging techniques, prediction-powered inference (Angelopoulos et al., 2023) and predict-then-debias estimators (Chen and Chen, 2000, Egami et al. (2024), Kluger et al. (2025), Lu et al. (2025)), also leverage validation data to correct for misclassification.[9] The former requires adjustment to a loss function, complicating estimation and limiting its applicability to M-estimation.[10] The latter is similar, but adjusts parameter estimates directly rather than loss functions, but outside of generalized linear models, requires either custom calculations or bootstrapping to compute confidence intervals. Relative to these approaches, multiple imputation offers an attractively easy-to-implement approach: estimation of the imputation model and generation of imputed datasets only needs to occur once, and from this initial effort, every summary statistic and model estimate in a research project can be calculated as normal without any modification to standard statistical software. Though other semi-parametric methods may offer better efficiency guarantees, multiple imputation's staying power in other literatures is rooted in its simplicity and wide applicability, with serious practical and logistical advantages for complex work involving many different estimators.

[9] Egami et al. (2024) refer to their method as design-based supervised learning, but the method's formulation is equivalent to the predict-then-debias approach.

[10] This can also result in a non-convex loss function, even when the loss function without the rectifier is convex (Ji et al., 2025).

Multiple Imputation

Problem Setup and Assumptions

I aim to estimate and conduct inference on a (potentially multidimensional) parameter β describing a probability distribution over random vectors (Z, X, Y) where Z and X are categorical and Y is continuous. This general setup allows Z to be either an outcome or independent variable. I observe a sample $\mathcal{S}$ of size N containing $(\tilde{Z}, X, Y)$, where $\tilde{Z}$ is a potentially misclassified predicted proxy for Z, e.g. a satellite-based prediction. For a subset $\mathcal{S}^* \subset \mathcal{S}$ of size N^*, I also observe the true value of Z; this subset forms the validation data. I denote by Z^+ the observed true value of Z for $i \in \mathcal{S}^*$ and missing otherwise, and by R an indicator equal to one if $i \in \mathcal{S}^*$. Let Z^- denote the unobserved true values of Z for $i \notin \mathcal{S}^*$. The observed dataset is therefore $(Z^+, \tilde{Z}, X, Y, R)$.

Analyzing misclassified data with accompanying validation data is analogous to analyzing data with missing values. Despite the analogy, two key differences distinguish the predicted data misclassification problem from a typical missing data problem. First, the extent of "missingness" is much higher: validation samples may cover only a small fraction of observations ($N^* << N$), whereas methods for missing data are generally designed assuming most values are observed. Second, I observe the proxy $\tilde{Z}$, which provides valuable information about the missing Z and simplifies imputation in practice. My strategy adapts multiple imputation methods from the missing data literature to exploit information in the misclassified proxy $\tilde{Z}$. Before continuing, I introduce the basic assumptions necessary to proceed with a multiple imputation based analysis. The first two assumptions concern sampling: the first deals with how the overall sample $\mathcal{S}$ is drawn from the population $\mathcal{P}$, while the second concerns how the validation sample $\mathcal{S}^*$ is drawn from $\mathcal{S}$.

Assumption 1 (Ignorable Sampling)**.** Consider the full population $\mathcal{P}$ of (Z, X, Y) of which our data are only a subset $\mathcal{S} \subset \mathcal{P}$. Let I be an indicator variable for whether an observation

in $\mathcal{P}$ is also in $\mathcal{S}$. The sampling procedure generating $\mathcal{S}$ is ignorable (Rubin, 1987, p. 50), meaning

$$P(I_i = 1 \mid Z_i, \tilde{Z}_i, X_i, Y_i, R_i) = P(I_i = 1 \mid \tilde{Z}_i, X_i, Y_i, R_i)$$

This is weaker than simple random sampling and allows for stratifying on observable information. Since sampling is assumed ignorable, I will be suppressed throughout the rest of my analysis.

Assumption 2 (Ignorable Validation Sampling)**.** The vectors $(Z_i, \tilde{Z}_i, X_i, Y_i, R_i)$ are i.i.d. with $P(R_i = 1 \mid Z_i, \tilde{Z}_i, X_i, Y_i) = \rho_N \pi(\tilde{Z}_i, X_i, Y_i)$, $\rho_N \in (0, 1]$, and $N\rho_N \to \infty$.

This assumption is again weaker than simple random sampling but requires that the probability of inclusion in the validation data does not depend on the unobserved true value of Z. In particular, this *does* allow for validation sampling based on observed values of $\tilde{Z}$, a likely feature of validation datasets created for the purpose of assessing the quality of a predicted data product.

Unfortunately, the ground truth data used to train the machine classifier that produced $\tilde{Z}$ violates this assumption. Because in-sample prediction accuracy is almost always higher than out-of-sample accuracy, observations where $Z = \tilde{Z}$ are more likely to be included in the training set, violating ignorability. Using such data in imputation would understate variance, since the training data over represents cases the classifier predicts well. In contrast, holdout samples used to evaluate classifer accuracy after training avoid this bias and offer a promising source of validation data.

Following the introduction of multiple imputation and my imputation model, I will discuss how violations of these assumptions affect estimation. I particularly focus on departures from ignorability in validation data sampling, given the potential availability of non-ignorable validation samples of convenience.

The important consequence of Assumptions 1 and 2 is that $P(Z^- \mid Z^+, \tilde{Z}, X, Y, R, I) =$

$P(Z^- \mid Z^+, \tilde{Z}, X, Y)$, i.e. when the validation sample is ignorable, there will not be systematic differences in the conditional probability of Z between observations inside and outside of the validation sample.[11] While consistent multiple imputation is still possible under certain violations of this assumption, implementation requires an additional auxiliary model relating observed variables to the validation sampling mechanism (Rubin, 1987, p. 212), which is beyond the scope of this paper.

Assumption 3 (Positivity in Validation Sampling)**.** The conditional probability of any given observation in the dataset $\mathcal{S}$ also appearing in the validation dataset $\mathcal{S}^*$ is positive, i.e.

$$P(R_i = 1 \mid \tilde{Z}_i, X_i, Y_i) > 0 \; \forall \; \tilde{Z}_i, X_i, Y_i$$

This is a basic identifiability assumption; if certain values of $(Z, \tilde{Z}, X, Y)$ are possible but cannot appear in the validation data, the imputation model would require out-of-sample extrapolation, even in the limit as $N^* \to \infty$.

Assumption 4 (Regularity of the Estimator)**.** The estimator of interest $\hat{\beta}$ (what would be directly estimated and consistent for β if complete and correctly classified data were available), depends on the data only through sample averages $\frac{1}{N}\sum_i f(Z_i, X_i, Y_i)$ of functions f in a class $\mathcal{F}$, and does so continuously, i.e. $\hat{\beta} \xrightarrow{p} \beta$ whenever these averages converge in probability to their population counterparts uniformly over $\mathcal{F}$. The class $\mathcal{F}$ satisfies a uniform law of large numbers for imputed samples via compactness, continuity, or bracketing conditions, with moment bounds that hold at every value of Z, not just the values in the support of the data's true distribution.

The first part of Assumption 4 is how consistency is proven for most estimators in econometrics (Newey and McFadden, 1994). It holds for M-estimators with compact parameters spaces or a convex maximand, which includes maximum likelihood, nonlinear least squares, and quantile

[11] This is analagous to the "missing at random" assumption in the missing data literature.

regression estimators; GMM estimators; continuous functions of sample means, including OLS and two-stage least squares estimators; and plug-in estimators of continuous functionals of the data distribution, such as quantiles. The misclassified variables can be an outcome, regressor, or even an instrument. The second part is where imputation differs: because the imputation model will have error, the imputed data may contain combinations of variables values that are not possible under the data's true distribution. Therefore, the functions $f \in \mathcal{F}$ must be well-behaved even outside the true data's support. When Z enters estimation only through indicators and through interactions of those indicators with functions of the other variables, this is implied by the integrability of those functions. Appendix 3 shows that under Assumptions 1-4 the multiple imputation estimator is consistent for β.

As the previous paragraph illustrates, Assumption 4 covers the vast majority of estimators used in applied econometrics. It is also a different than assuming asymptotic linearity, as is common in other semiparametric methods for handling measurement error (Dell, 2025; Kluger et al., 2025). Though there is substantial overlap in what each of these assumptions covers, neither is subsumed by the other. Maximum score estimators, for example, are covered by Assumption 4 but are not asymptotically linear. Semiparametric two-stage estimators with a nonparametric first stage can be asymptotically linear, but fail Assumption 4. An even wider class of estimators fail both assumptions, e.g. local kernel estimators or penalized estimators like LASSO. To be clear, these other estimators *may* be consistent with multiple imputation, but proving so would require a different argument than I provide.

Multiple Imputation Basics

Although multiple imputation has strong frequentist justification, it is most easily understood in a Bayesian framework. I therefore describe the approach in terms of estimating a posterior for β.

With fully observed data, the posterior distribution of β with complete data is $\mathrm{P}(\beta \mid Z, X, Y)$,

which by assumption is equivalent to $\mathrm{P}(\beta \mid Z, \tilde{Z}, X, Y)$ since the misclassified proxy $\tilde{Z}$ adds no information once Z is known. Instead, Z is only partially observed, so the actual posterior distribution is $\mathrm{P}(\beta \mid Z^+, \tilde{Z}, X, Y)$. These two posteriors are related by

$$\mathrm{P}(\beta \mid Z^+, \tilde{Z}, X, Y) = \int \mathrm{P}(\beta \mid Z, X, Y) \times \mathrm{P}(Z^- \mid Z^+, \tilde{Z}, X, Y)\, dZ^-$$

Following Rubin (1987, p. 82), this formulation implies that draws from $P(Z^- \mid Z^+, \tilde{Z}, X, Y)$ can be used to simulate the actual posterior of β, since Z^+ and Z^- together determine Z. The following section will detail the precise imputation model I use to estimate this conditional distribution. Generating imputations of Z^- from this conditional distribution fills in the unobserved cases Z^-, after which the complete data estimator $\hat{\beta}$ can be applied without modification. Repeating this process M times and recording the estimate $\hat{\beta^m}$ and variance estimate U^m each time yields $\{\hat{\beta}^m, U^m\}_{m=1}^M$.

Again following Rubin (1987, p. 84), let $\hat{\beta}$ be the mean of the posterior distribution of β with complete data and U its associated variance-covariance matrix. Then by the law of iterated expectations,

$$\begin{aligned}
\mathbb{E}(\beta \mid Z^+, \tilde{Z}, X, Y) &= \mathbb{E}(\mathbb{E}(\beta \mid Z, X, Y) \mid Z^+, \tilde{Z}, X, Y) \\
&= \mathbb{E}(\hat{\beta} \mid Z^+, \tilde{Z}, X, Y)
\end{aligned}$$

and by the law of total conditional variance,

$$\begin{aligned}
\mathrm{Var}(\beta \mid Z^+, \tilde{Z}, X, Y) &= \mathbb{E}(\mathrm{Var}(\beta \mid Z, X, Y) \mid Z^+, \tilde{Z}, X, Y) \\
&\quad + \mathrm{Var}(\mathbb{E}(\beta \mid Z, X, Y) \mid Z^+, \tilde{Z}, X, Y) \\
&= \mathbb{E}(U \mid Z^+, \tilde{Z}, X, Y) + \mathrm{Var}(\hat{\beta} \mid Z^+, \tilde{Z}, X, Y)
\end{aligned}$$

For large M, this suggests the following estimators (Rubin, 1987, p. 85):

$$\hat{\beta}_\infty \equiv \lim_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} \hat{\beta^m} = \mathbb{E}(\hat{\beta} \mid Z^+, \tilde{Z}, X, Y)$$

$$U_\infty \equiv \lim_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} U^m = \mathbb{E}(U \mid Z^+, \tilde{Z}, X, Y)$$

$$B_\infty \equiv \lim_{M\to\infty} \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\beta^m} - \hat{\beta}_\infty\right)\left(\hat{\beta^m} - \hat{\beta}_\infty\right)' = \mathrm{Var}(\hat{\beta} \mid Z^+, \tilde{Z}, X, Y)$$

These satisfy $\mathbb{E}(\beta \mid Z^+, \tilde{Z}, X, Y) = \hat{\beta}_\infty$ and $\mathrm{Var}(\beta \mid Z^+, \tilde{Z}, X, Y) = U_\infty + B_\infty$, implying $(\hat{\beta}_\infty - \beta \mid Z^+, \tilde{Z}, X, Y) \sim \mathcal{N}(0, U_\infty + B_\infty)$.

In practice, M is finite and may even be quite small depending on the computational burden of parameter estimation. Rubin (1987) assesses the small-M case and suggests using $T_M = U_M + (1 + \frac{1}{M})B_M$ with U_M and B_M to estimate the total variance.

For inference, multiple imputation estimators use t-distributions, but require calculating an adjusted degrees of freedom to account for the imputation. For general use, where v_{com} is what the degrees of freedom would be with complete truth data, Barnard and Rubin (1999) and Reiter and Raghunathan (2007) recommend the following formula for its satisfactory frequentist coverage:

$$v_m = \left[\left((M-1)\left(1 + \frac{U_M}{\left(1+\frac{1}{M}\right)B_M}\right)^2\right)^{-1} + \left(v_{com}\frac{U_M}{T_M}\frac{v_{com}+1}{v_{com}+3}\right)^{-1}\right]^{-1}$$

With this foundation in place, the next task is to specify an imputation model for $\mathrm{P}(Z^- \mid Z^+, \tilde{Z}, X, Y)$. Because the dependence of misclassification on Z, X, and Y is uncertain and imputation model misspecification could cause bias, I turn next to a flexible nonparametric approach for generating draws of the unobserved Z values.

A Bayesian Nonparametric Imputation Model

The key step in implementing multiple imputation is obtaining a sensible estimate of the conditional distribution $\mathrm{P}(Z^- \mid Z^+, \tilde{Z}, X, Y)$, which governs the unobserved truth values for the misclassified variable. Estimating this probability mass function allows me to draw multiple sets of imputed Z values and proceed with inference as described in the previous section.

Since Z is categorical, this conditional distribution is itself categorical, with probabilities for each category dependent on $Z^+, \tilde{Z}, X$, and Y. However, imputing from a single best-fit distribution would treat the estimated probabilities as known with certainty, ignoring uncertainty about the distribution itself.[12] Instead, I seek a *posterior* distribution over $\mathrm{P}(Z^- \mid Z^+, \tilde{Z}, X, Y)$, informed by a minimally informative prior and the validation data. This posterior allows me to first draw a probability vector for each unvalidated observation and then draw one realization of z_i from that vector for each imputation m. As Rubin (1987) emphasizes, this approach propagates uncertainty in the imputation process and yields valid confidence intervals.

Many choices of imputation model are possible. When misclassification is known to take a specific form or to be independent of some covariates, bespoke models can deliver tighter inference. Here, my goal is to construct a general purpose imputation model that accomodates mixed continuous and categorical data, avoids parametric misspecification, and is parsimonious enough to be feasible to estimate.[13] A promising approach satisfying these criteria is the Hierarchically Coupled Mixture Model with Local Dependence (HCMM-LD) first introduced by Murray and Reiter (2016). This Bayesian nonparametric model flexibility represents the joint distribution of mixed continuous and discrete variables, growing more complex as data

[12] Structural approaches instead incorporate uncertainty over misclassification probabilities by embedding the parameters governing misclassification into the likelihood function and estimating them directly.

[13] Even under these criteria, there are promising alternatives to the model I use. As one example, sequential regression trees (Burgette and Reiter, 2010) have shown promise in empirical applications (Akande et al., 2017). The consistency proofs in Appendix 3 are not specific to this particular imputation model, and would work for any imputation model that converges in probability to the true data distribution in L^1 distance.

demand while remaining parsimonious.

Model Structure

Building on the success of mixture models methods for purely categorical or continuous datasets (Murray, 2018), the HCMM-LD couples two lower-level mixture models, one for categorical variables and one for continous variables conditional on the categorical variables, through a third, top-level mixture. This hierarchical coupling maintains flexibility while preventing overparameterization by allowing conditional independence between the lower-level components while preserving dependence through the higher-level one. The model is estimated using the fully observed validation data.[14]

Expositionally, it helps to split the data into categorical variables and continuous variables, lumping what I have been calling Z and $\tilde{Z}$ in with what I have been calling X. Let $X_i = (Z_i, \tilde{Z}_i, X_{i1}, \ldots, X_{ip-2})'$ be a vector of p categorical variables, with the first and second elements equal to the true and potentially misclassified versions of the variable we are imputing for, respectively. Let $Y_i = (Y_{i1}, \ldots, Y_{iq})'$ be a vector of continuous variables in $\mathbb{R}^q$.

Each individual is assigned to a categorical component $s_i \in \{1, \ldots, k^{(\mathcal{X})}\}$ and a continuous component $r_i \in \{1, \ldots, k^{(\mathcal{Y})}\}$. Conditional on memberships in these components, densities for X_i and Y_i are given by

$$\mathbb{P}(X_i = x_i | S_i = s) = \prod_{j=1}^{p} \psi_{sx_{ij}}^{(j)}$$

$$(Y_i = y_i | R_i = r, X_i = x_i) \sim \mathcal{N}(D(x_i)B_r, \Sigma_r)$$

[14]This differs from Murray and Reiter (2016), who estimate the HCMM-LD using all data and impute missing values within the Gibbs Sampler. This is due to their focus on situations in which many variables have a small degree of missing information, in which case using all data can be valuable, since an observation with missing data for one variable still contains information about the covariance of other variables that may be missing in other observations. This is not a concern for my intended application. The non-validated data do not contribute much information about the relationship of Z to other variables, and excluding them saves significant computational resources in the Gibbs Sampler.

A higher-level component $z_i \in \{1, \ldots, k^{(\mathcal{Z})}\}$[15] links the lower-level components,

$$\mathbb{P}(S_i = s, R_i = r | Z_i = z) = \phi_{zs}^{(\mathcal{X})} \phi_{zr}^{(\mathcal{Y})}$$

with membership in the high-level component drawn from a categorical distribution with probabilities

$$\mathbb{P}(Z_i = z) = \lambda_z$$

Though categorical variables are mutually independent within a component, dependence among categorical variables arises through shared component membership; dependence between categorical and continuous variables is captured directly through the regression structure $D(x)B_r$ (with $D(x)$ being the design matrix of x and B_r being parameter estimates), with additional higher-order interactions captured via the high-level component.[16] Dependence between continuous variables is also modeled directly through Σ_r. As B_r and Σ_r are specific to each continuous component, these parameters capture *local* rather than global dependence.

This structure enables a great deal of flexibility from a parsimonious setup. The marginal distribution of X takes a latent class form,

$$\mathbb{P}(X_i = x_i) = \sum_{s=1}^{k^{(\mathcal{X})}} \left(\sum_{z=1}^{k^{(\mathcal{Z})}} \lambda_z \phi_{zs}^{(\mathcal{X})} \right) \prod_{j=1}^{p} \psi_{sx_{ij}}^{(j)}$$

which reduces to the model of Si and Reiter (2013) or (with an unrestricted number of components) Dunson and Xing (2009), which is consistent for *any* multivariate categorical distribution, a consequence of the fact that any such distribution can be expressed as a finite mixture of product multinomial distributions.[17] The marginal distribution of Y is a mixture

[15] This Z is not to be confused with my earlier notation for the true categorical variable, which is now a component of X. I use Z to refer to the latent top-level mixture component to maintain notation consistency with Murray and Reiter (2016).

[16] $D(x)B_r$ simply alleviates the proliferation of components.

[17] Of course, this applies only to multivariate categorical distributions with finite variables and finite categories. A "product multinomial" distribution is a joint distribution of independent categorical variables. See Dunson and Xing (2009) for proofs.

of multivariate normals, which is itself consistent for any twice continuously differentiable smooth density (Ghosal and van der Vaart, 2007) or even for non-smooth densities satisfying mild regularity conditions (Wu and Ghosal, 2010).

Marginalizing over all the components yields the joint density:

$$f(X_i, Y_i) = \sum_{z=1}^{k^{(\mathcal{Z})}} \lambda_z \left(\sum_{r=1}^{k^{(\mathcal{Y})}} \phi_{zr}^{(\mathcal{Y})} \mathcal{N}(Y_i; D(x_i)B_r, \Sigma_r) \right) \left(\sum_{s=1}^{k^{(\mathcal{X})}} \phi_{zs}^{(\mathcal{X})} \prod_{j=1}^{p} \psi_{sx_{ij}}^{(j)} \right)$$

This structure straightforwardly enables drawing conditional probabilities for any categorical variable:

$$\mathbb{P}(X_{ij} = c_j | X_{i/j} = x_{i/j}, Y_i) \propto \sum_{z=1}^{k^{(\mathcal{Z})}} \lambda_z \left(\sum_{r=1}^{k^{(\mathcal{Y})}} (\phi_{zr}^{(\mathcal{Y})} \mathcal{N}(Y_i; D(\tilde{x}_i)B_r, \Sigma_r)) \right) \left(\sum_{s=1}^{k^{(\mathcal{X})}} \phi_{zs}^{(\mathcal{X})} \prod_{l=1}^{p} \psi_{s\tilde{x}_{il}}^{l} \right)$$

where $X_{i/j}$ is the X_i-vector with the jth element removed and $\tilde{x}_i$ is the vector with entries equal to x_i except that the jth element is set to c_j. Appendix 2 proves that both the joint and conditional distributions estimated in this model converge in L_1 to their true counterparts for any marginal distribution of the categorical variables and subject to only mild regularity conditions on the continuous variables.

Estimation via Gibbs Sampling

The imputation model is estimated using a Gibbs Sampler. As an MCMC method, the Gibbs Sampler cycles through blocks of parameters and draws their values from the model's distributions of those parameters conditional on the current value of all other parameters. After convergence, each imputation can come from a single iteration of the Gibbs Sampler, with successive imputations calculated from distinct iterations.[18] This captures uncertainty in the categorical probabilities by using separate draws for each imputation. Full algorithmic details and priors are described in detail in Appendix 1.

[18] In simulation and applications, I allow multiple iterations between imputations to account for the dependence between neighboring draws in the Markov chain, though this should not matter if the model has appropriately converged.

Given the central importance of the hierarchical component structure, it is worth paying some attention to the priors for component membership and how they are updated. The priors for the λs and ϕs that determine component membership are distributions defined using the stick-breaking representation of the Dirichlet process (Sethuraman, 1994):

$$\phi_{zs}^{(\mathcal{X})} = \xi_{zs}^{(\mathcal{X})} \prod_{l<s}(1-\xi_{zl}^{(\mathcal{X})}), \quad \phi_{zr}^{(\mathcal{Y})} = \xi_{zr}^{(\mathcal{Y})} \prod_{l<r}(1-\xi_{zl}^{(\mathcal{Y})}), \quad \lambda_z = \xi_z^{(\mathcal{Z})} \prod_{l<z}(1-\xi_l^{(\mathcal{Z})})$$

where

$$\xi_{zs}^{(\mathcal{X})} \sim \text{Beta}(1, \beta^{(\mathcal{X})}), \quad \xi_{zr}^{(\mathcal{Y})} \sim \text{Beta}(1, \beta^{(\mathcal{Y})}), \quad \xi_z^{(\mathcal{Z})} \sim \text{Beta}(1, \alpha)$$

except for $\xi_{zk^{(\mathcal{X})}}^{(\mathcal{X})}$, $\xi_{zk^{(\mathcal{Y})}}^{(\mathcal{Y})}$, and $\xi_{k^{(\mathcal{Z})}}^{(\mathcal{Z})}$, which are set equal to one, truncating the stick-breaking process. Truncation substantially eases computation with negligible loss in flexibility (Ishwaran and James, 2001), but requires checking that, at convergence, some components remain empty or near empty. If all components (in any dimension) are occupied by more than a trivial number of observations, that is evidence of excessive truncation and estimation should begin again with $k^{(\cdot)}$ set higher in the relevant dimension.

This construction permits data-driven complexity: the concentration parameters $\beta^{(\mathcal{X})}$, $\beta^{(\mathcal{Y})}$, and α evolve with the number of occupied components. In each iteration, each observation is assigned to one of each of the three components, with components that accurately fit that observation receiving higher probability weight. When none of the existing components fit well, that leaves greater probability mass in empty components, which can lead that observation to initializing a new component to be considered by all observations in future runs. Abstracting from the model's probabilistic nature, the estimation procedure keeps the model's complexity static when existing observations are well described by some combination of existing components, and automatically increases the complexity otherwise.

Computational Feasibility

The proposed implementation of the HCMM-LD is computationally feasible even for extremely large predicted datasets. The computationally intensive step, which is estimating the imputation model, occurs only on the comparatively small validation dataset (though the consistency proof does require $N^* \to \infty$, there is no prohibition on $N^*/N \to 0$). The complexity of this computation scales roughly linearly with sample size, though complexity can rise quickly with the number of continuous variables due to the covariance matrices (Murray and Reiter, 2016). Crucially, estimation of the imputation model occurs only once and only until convergence. Once fitted, the model can generate imputations for the large predicted dataset with minimal additional computation, and this step can be parallelized to handle truly massive datasets efficiently.

Incomplete Covariates in the Validation Data

As discussed above, implementing multiple imputation requires that all covariates in the analysis model (that is, all covariates you would use to estimate $\hat{\beta}$ with complete data) be observed within the validation data, so their joint dependence with the true distribution of Z can be modeled through $\mathrm{P}(Z^- \mid Z^+, \tilde{Z}, X, Y)$. In practice, however, a validation sample $\mathcal{S}^*$ containing Z^+ and $\tilde{Z}$ may lack some or all of the other model covariates.[19] In such settings, multiple imputation estimation remains feasible if the missing covariates are conditionally independent of Z given the observed information. Let X_a and Y_a be the observed covariates and X_b and Y_b the unobserved covariates in the validation sample. If $Z^- \perp\!\!\!\perp X_b, Y_b \mid Z^+, \tilde{Z}, X_a, Y_a$, then $\mathrm{P}(Z^- \mid Z^+, \tilde{Z}, X, Y) = \mathrm{P}(Z^- \mid Z^+, \tilde{Z}, X_a, Y_a)$, which can be estimated using the imputation model. That this is possible is another advantage of the multiple imputation method over predication-powered inference or predict-then-debias approaches.

[19] In this case, $\mathcal{S}^* \not\subset \mathcal{S}$; rather S and S^* are ignorable samples from the same population $\mathcal{P}$. This is an alternative scenario to that assumed throughout the rest of the paper, in which $\mathcal{S}^* \subset \mathcal{S}$.

This issue is particularly likely to arise in randomized program evaluations. Researchers may have comprehensive predicted data and a sample of validation data drawn from a disjoint or minimally overlapping portion of the same population over which treatment assignment is randomized. When Z is an independent variable in the model of interest $Z^- \perp\!\!\!\perp X_b \mid Z^+, \tilde{Z}, X_a, Y$ is clearly a safe assumption when X_b is randomized treatment. However, if Z is the outcome, this assumption no longer holds: when treatment potentially affects Z, it is generally implausible to assume that misclassification error in the predicted data is independent of treatment, as the predictions themselves are post-treatment quantities. This concern does not arise if treatment data are observed in the validation sample, since dependence can then be explicitly modeled. When this is not available, Rambachan et al. (2025) offer an approach for estimation tailored to this specific setting.

Non-Ignorable Validation Data

A key assumption underpinning the multiple imputation procedure is that the validation sampling mechanism depends only on observed variables. Ignorability ensures that the imputation model estimated on the validation sample can be validly extrapolated to the full dataset. When this assumption fails, the validation data become a selected sample, and the imputation model inherits that selection bias.

In practice, however, validation data that does not arise from formal randomization are often the only validation data available. A key question is how sensitive multiple imputation is to a departure from this assumption, and whether the formally inconsistent estimates that arise from applying multiple imputation with a non-ignorable validation sample should be expected to mitigate bias relative to naively using predicted data.

The impact of non-ignorability depends on the relationship between the selection mechanism and the true underlying variable and the sensitivity of the model being estimated. Suppose W is an unobserved variable omitted from the imputation model and that $P(R) = f(W)$.

If W is strongly correlated with Z (conditional on $\tilde{Z}$, X, and Y), bias can be substantial. Under weaker correlations, however, multiple imputation often remains beneficial: it reduces bias relative to naive estimation even when ignorability is moderately violated. Simulation and empirical results shown in the following sections will illustrate this robustness property, mirroring similar findings by Proctor et al. (2023) and others in the multiple imputation literature.

When the source of non-ignorability is known, it can often be addressed directly by including relevant covariates in the imputation model so long as the separate positivity condition still holds. For example, it may be significantly easier to collect validation data for a satellite dataset by sampling locations adjacent to accessible roads. Because road proximity might correlate with whatever land characteristic is measured in the satellite data even after conditioning on other covariates, this constitutes a non-ignorable sampling design. However, including road proximity as an auxiliary variable in the imputation model would restore ignorability, so long as all levels of road proximity have nonzero sampling probability. This does not require any adjustment to the ultimate model of interest—it pertains only to the imputation step.

Misclassification in the Validation Sample

Even "validation" data are rarely perfect. Ground truth measurements may contain their own modest degree of misclassification, reflecting the inherent limitation of any data collection process. In this setting, multiple imputation recovers the parameter estimates that would be obtained if *all* observations were measured with the same accuracy as the validation data, not the estimates that would arise under truly error-free measurement.

Error in the multiple imputation estimator can be decomposed as $\text{Error}(\hat{\beta}^{MI}) = (\hat{\beta}^{MI} - \hat{\beta}) + (\hat{\beta} - \beta)$, where $\hat{\beta}$ is the infeasible value of the estimator that would be recovered if data of the same quality as the validation data were available for the entire dataset and used

in estimation. The first of these terms vanishes asymptotically: the multiple imputation estimator converges to the same limit as the complete data estimator.[20] Therefore, multiple imputation neither resolves nor amplifies bias present in the validation data; it merely extends the level of accuracy in that data to the full dataset.

Special Considerations for Panel Data

Econometric applications of panel data often include large sets of fixed effects, estimated as nuisance parameters. This poses a distinct challenge for imputation: fixed effects must be represented within the imputation model, or the dependence between observations sharing the same fixed effect will be lost in the imputed data. This requirement is often described as *congeniality* between the imputation model and the model of interest, i.e., the imputation model must allow for at least as rich a dependence structure as the model of interest (Meng, 1994).

The severity of this problem depends on the granularity of the fixed effects. In settings where the fixed effects correspond to coarser grouping, such as states or years in individual-level panels, incorporating them into the imputation model is straightforward. Each grouping variable can be supplied as a categorical variable in the imputation model (Rubin, 1996).[21] The design matrices for such variables are relatively compact, so computation remains tractable. This will generally suffice for models with misclassified outcome data, since the incidental parameters problem typically precludes the use of fine-grained fixed effects in nonlinear models anyway.

When fixed effects are high-dimensional, however, the curse of dimensionality becomes more acute. While there is nothing *mathematically* preventing the imputation model from

[20] Derivation of this fact is given in Appendix 3.

[21] However, since this imputation model is fully nonparametric, inclusion of group and time fixed effects will automatically cover any other variables that only vary at the group, time, *or group* $\times$ *time* level. In a standard two-way fixed effect design, this would mean treatment can be excluded from the imputation model, as it only varies at the group $\times$ time level. In fact, beyond the fixed effects themselves, the only variables necessary to include would be those varying over time at the sub-group level.

accommodating thousands of fixed effects, estimation becomes unstable: sparse indicators contribute little information, runtime grows rapidly, and the Gibbs Sampler may fail to converge. In such cases, it is preferable to conduct imputation on a *wide* version of the dataset, where each time period's observation forms a separate variable with an individual's record, rather than the *long* version commonly used for estimation. If the validation data do not include complete time series for each individual, the Gibbs Sampler can be modified to draw the missing elements of the validation data in each iteration. This strategy would not alter the theoretical properties of the multiple imputation estimator, nor would it affect how downstream estimation occurs, since the imputed datasets can easily be converted back to long form.

Simulations

To evaluate the performance of multiple imputation using the HCMM-LD model, I conduct a series of simulations comparing four estimation strategies:

1. Complete Data Estimate: These estimates use the true value of the misclassified variable, Z. Because most of Z would be unobserved in an applied setting, this estimator is infeasible in practice but serves as an ideal benchmark: it isolates the amount of variation in parameter estimates attributable to sampling variability alone.

2. Naive Estimate: These estimates use the potentially misclassified version of the variable, $\tilde{Z}$, without correction. This corresponds to the standard practice of estimating models directly from imperfect predicted data, leading to bias and inconsistency.

3. Validation Only Estimate: These estimates are based solely on the validation sample. Since multiple imputation requires such a sample, it will often be possible to estimate the model of interest using it alone, though at the cost of reduced statistical power.[22]

[22]When Z is the dependent variable and the validation data sampling is stratified on $\tilde{Z}$, however, multiple imputation becomes more than a simple power improvement, since the validation data are selected on (a potentially misclassified version of) the dependent variable. This still satisfies the ignorability assumption

Except insofar as the reduced sample size makes asymptotic approximations unreliable, the validation only estimates should share the same expectation as the complete data estimates, differing primarily by having much wider confidence intervals. This comparison highlights the benefit of multiple imputation in combining information from the large, imperfect dataset to achieve narrower intervals without sacrificing consistency.

4. Multiple Imputation Estimate: These estimates use imputations generated from the HCMM-LD model. Details specific to each simulation design are provided in the sections that follow.

Conditionally Random Misclassification in the Outcome of a Simple Binary Logit

The initial simulation tests the performance of the multiple imputation estimator in the presence of conditionally random misclassification in the outcome of a binary logit. Specifically, the binary outcome y is misclassified with probability $\mathbb{P}(y \neq \tilde{y}|x) = 0.2\frac{e^x}{1+e^x}$, where $x \sim \mathcal{N}(0,1)$ is the sole covariate (the dependence here is not strictly necessary to demonstrate the point: even completely random misclassification in $\tilde{y}$ would induce bias resolved by multiple imputation). This specification yields an overall misclassification rate of 0.1, with the probability of misclassification increasing in x. Consequently, true values of $y = 1$ are more likely to be misclassified as $\tilde{y} = 0$ than vice versa.

The simulation uses the following data generating process:

$$x \sim \mathcal{N}(0,1), \quad e \sim \mathcal{N}(0,1), \quad y = \mathbf{1}(x + e > 0)$$

$$N = 5000, \quad N^* = 500, \quad \mathbb{P}(y \neq \tilde{y} \mid x) = 0.2\frac{e^x}{1+e^x}$$

Figure 2 presents distributions of parameter estimates for each of the four estimation strategies

required for multiple imputation.

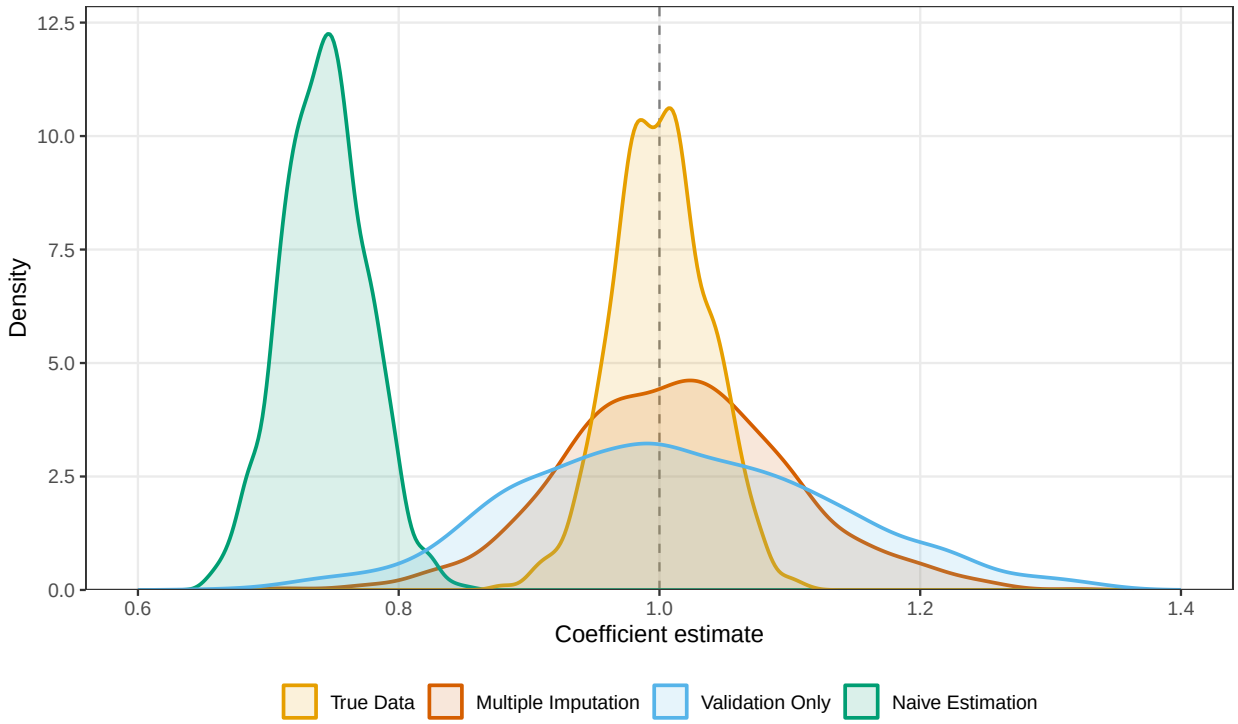


Figure 2: Density of Parameter Estimates from Simple Logit DGP with Dependent Misclassification

across 1000 simulated datasets. First and foremost, the naive estimate produces a badly biased and low variance result, with an empirical distribution of estimates that does not contain the true coefficient value of 1. The multiple imputation and validation both lack any sign of bias, and the multiple imputation estimate has a tighter distribution.

The empirical coverage of all non-naive estimates is close to nominal, with the multiple imputation estimate at 0.953 and the validation only estimate at 0.952.

Influence of the Validation Share

Given the importance of validation data to implementing the multiple imputation method, it is necessary to ask how much validation data is needed for the method to work well, or, at the very least, how much is needed for the multiple imputation method to outperform the naive estimate. In general, many factors influence the answer to this question, including the severity of misclassification, the informativeness of other covariates in predicting misclassification, and the structure of the ultimate economic model to be estimated. Still, in the interest of shedding some light on these questions, this simulation repeats the previous simulation at various validation shares and overall sample sizes.

Figure 3 presents the mean and empirical 95% interval of estimates for both the multiple imputation and validation only estimators across sample sizes $N \in \{1000, 5000, 10000\}$ and validation shares $N^*/N \in \{0.01, 0.05, 0.1, 0.2, 0.3\}$.[23] Also reported are the true data estimate and native estimate, which are not shown with intervals because they do not involve using a share of validation data.

Figure 3 demonstrates that for this data generating process, multiple imputation is approximately unbiased even at very low validation shares, that multiple imputation always outperforms the validation only estimator, and that most of the benefit of multiple imputation can be achieved at a relatively low amount of validation data. Moreover, the absolute

[23] Results for $N = 1000$ and $N^*/N = 0.01$ are excluded, as this leaves only 10 validated observations.

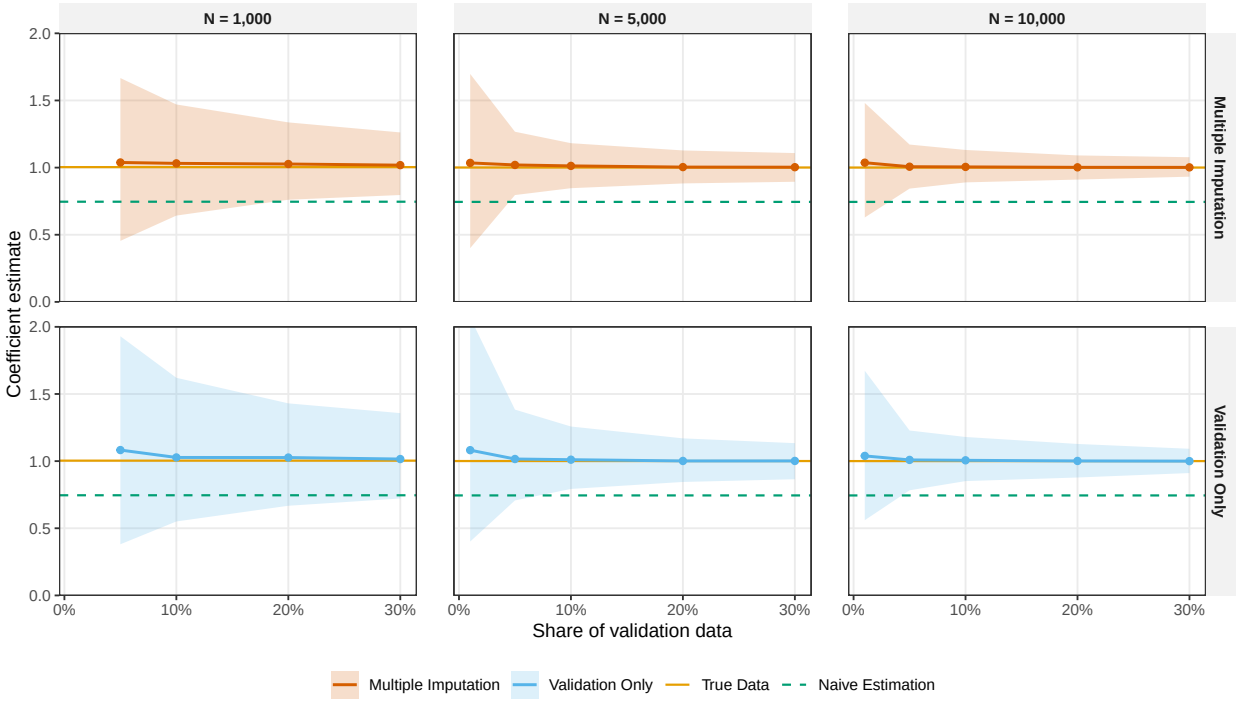


Figure 3: Influence of the Validation Share on Multiple Imputation's Performance

amount of validation data seems to matter more than its share: the performance of multiple imputation with $N = 5000$ and $N^*/N = 0.1$ is roughly equivalent to its performance at $N = 10000$ and $N^*/N = 0.05$.

Non-ignorable Validation Sampling

All previous simulations used validation data that perfectly satisfied the assumption of ignorable validation sampling. In real applications, validation samples of convenience may not satisfy this condition. These simulations are designed to assess whether multiple imputation remains worthwhile when only non-ignorable validation data are available. This is a setting where multiple imputation is theoretically biased and inconsistent. The goal is to assess the magnitude of this bias in comparison to a naive estimate that applies no correction, which is itself also biased and inconsistent.

For these simulations, I use the following data generating process:

$$x \sim \mathcal{N}(0,1), \quad e \sim \mathcal{N}(0,1), \quad y = \mathbf{1}(x + e > 0)$$

$$N = 5000, \quad N^* = 500, \quad z \sim \text{Bernoulli}(0.25), \quad \mathbb{P}(y \neq \tilde{y} \mid z) = \begin{cases} 0.25 & z = 1 \\ 0.05 & z = 0 \end{cases}$$

This setup largely mirrors the original simulation, with the addition of a set of "hard to classify" observations. These observations, for which $z = 1$, are more likely to be misclassified than those for which $z = 0$. The setup is such that the overall misclassification rate remains 0.1. Note that misclassification here is completely independent of x.

The simulation samples varying shares of its validation data from "hard to classify" observations where $z = 1$. When that share is 25%, the validation sampling mechanism is ignorable (because the probability of being sampled is the same conditional on either $z = 1$ or $z = 0$). At all other shares, the validation sampling mechanism is non-ignorable and should result in bias.

Figure 4 plots the means and empirical 95% intervals for each possible share of "hard to classify" units in $\{0.025, 0.5, 0.1, 0.15, 0.2, 0.25\}$. As expected, the multiple imputation estimate exhibits a bias whenever this share is $\neq .25$ and thus the validation sampling mechanism is non-ignorable. However, even in the extreme 0.025 share case, the empirical 95% interval of the multiple imputation estimates is within the 95% interval of the validation only estimates and outperforms the naive estimate. This accords with received wisdom from decades of practioners' experience that multiple imputation tends to improve results (in terms of reducing bias relative to native estimation) even under moderate violations of its ignorability assumptions.

Figure 5 plots the empirical coverage of each estimator across the various sampling shares. While the coverage of the multiple imputation estimate is noticeably below the nominal 0.95

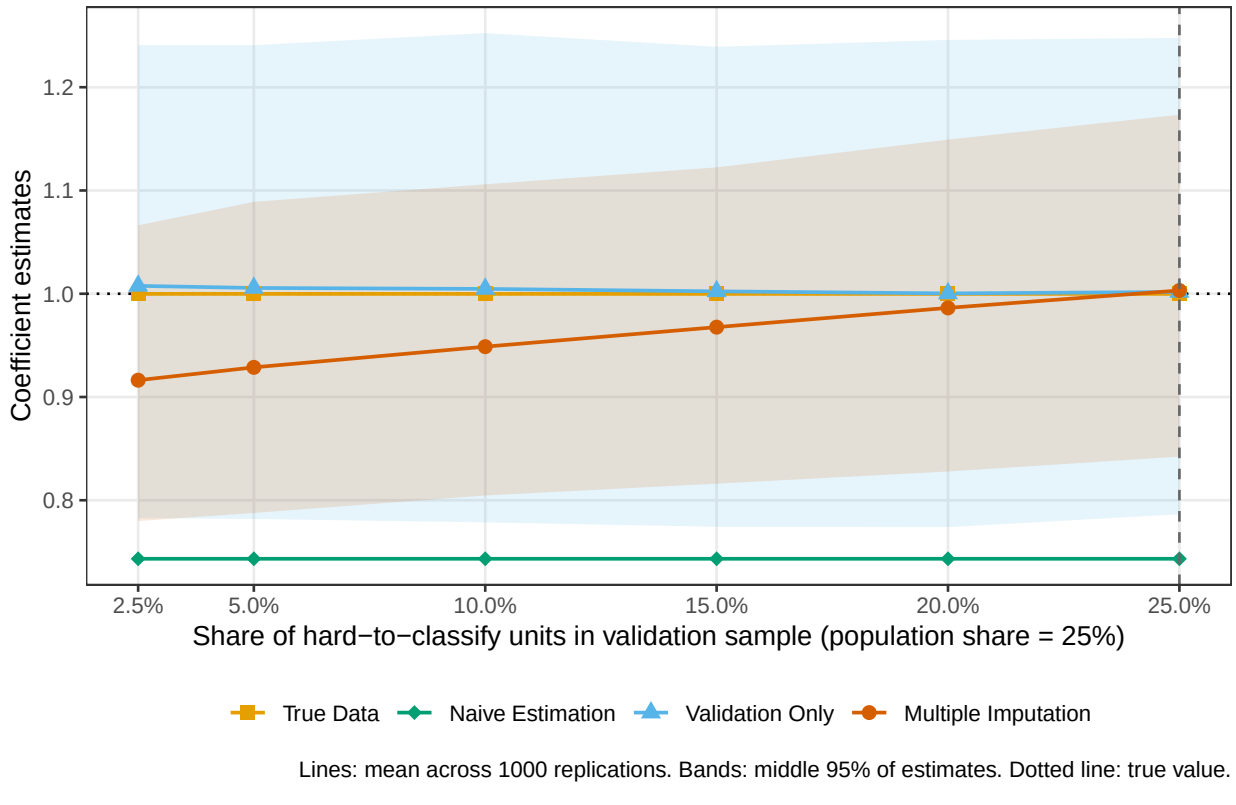


Figure 4: Results from Non-ignorable Sampling Simulations

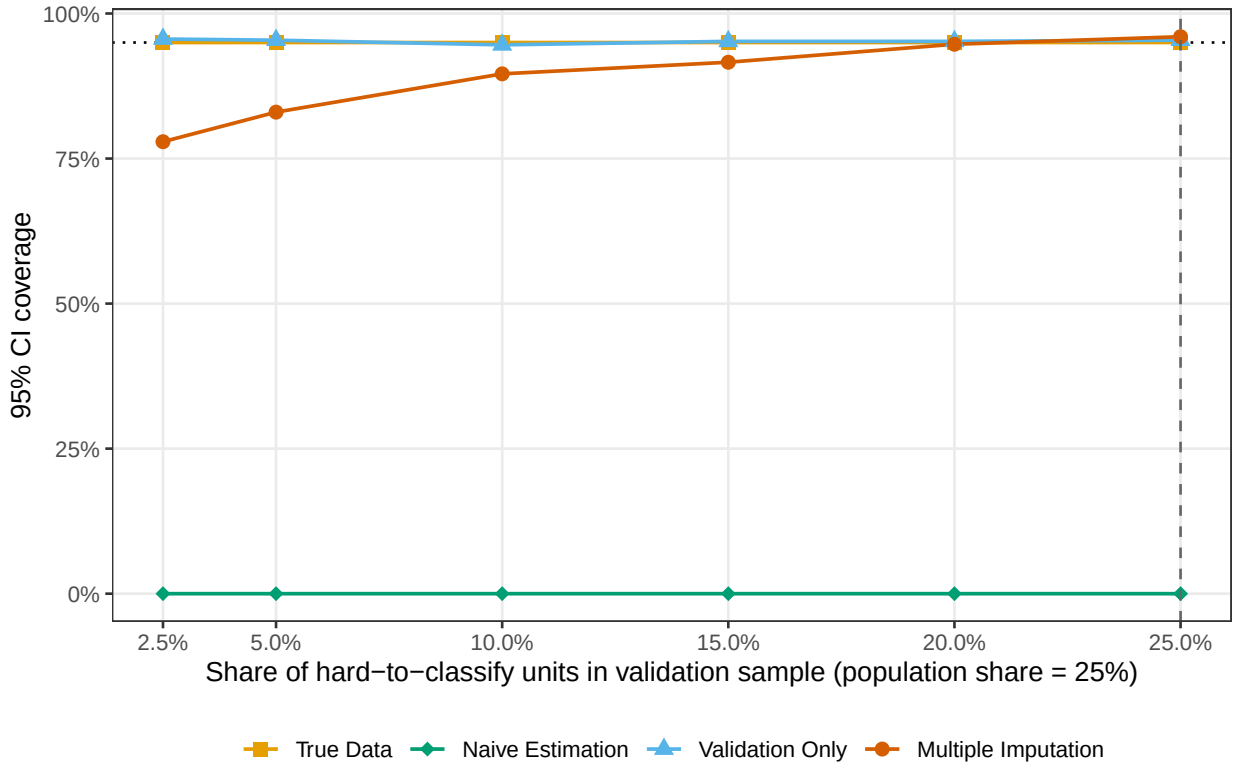


Figure 5: Results from Non-ignorable Sampling Simulations

for more severe non-ignorability, it gets quite close to nominal coverage in less severe cases and always outperforms the complete lack of overlap in the naive estimate. While the validation only estimate always maintains near-nominal coverage, it achieves this by having much wider confidence intervals. (In a case where the unobserved covariate was also correlated with the outcome, the validation only estimate would be biased and below nominal coverage as well.)

Of course, even more extreme forms of selection into the validation sample would change these results, and could potentially even make the non-ignorable multiple imputation estimate worse than the naive estimates. Still, the violation of ignorability in this simulation is non-trivial, and the results suggest that the multiple imputation approach remains reasonably robust to deviations from this assumption.

Taken together, the simulation results demonstrate that the HCMM-LD–based multiple imputation approach consistently performs well across a range of data-generating processes. Even when the probability of misclassification depends on covariates or when the validation sample violates ignorability, the method substantially reduces bias relative to naive estimators and maintains near-nominal coverage. These findings provide strong motivation to evaluate the approach in a real-world setting, where data limitations and complex dependence structures pose additional challenges.

Empirical Application: Demand for Short-run Water in Kern County

Satellite-based crop datasets, such as the USDA's Cropland Data Layer (CDL), are now ubitquitous in applied agricultural and environmental economics. These predicted datasets enable large-scale studies of crop choice and resource use but are known to contain classification errors that can bias downstream econometric analyses. I evaluate the accuracy of the CDL for Kern County, California, a major agricultural region at the southern end of the Central

Valley, and show how misclassification in the CDL affects estimated relationships between water availability and crop choice.

Kern County is an ideal setting for this exercise. It is one of the most agriculturally productive counties in the United States, ranking third nationally in total crop sales in the most recent USDA Census of Agriculture. Due in large part to the county's heavy emphasis on diverse specialty crops, the CDL is known to have accuracy issues in this region (Espinoza et al., 2023). At the same time, Kern County's Deparment of Agricultural and Measurement Standards provides comprehensive, field-level ground truth data, a rare resource among U.S. counties.[24] Created in partnership with local growers, largely in support of a pesticide use reporting program, this dataset enables direct comparison between true crop choices and CDL predictions. Figure 6 shows a map of Kern County fields and high-level crop categories in 2022.

Data Sources and Accuracy Assessment

The CDL is generated from multiple remote sensing sources, including NASA/USGS Landsat imagery (USDA NASS, 2024). Raw data spectral readings from these satellites are georeferenced and calibrated for terrain, atmospheric conditions, solar angles, and sensor characteristics. The quality of these adjustments depend on the quality of other inputs, including the model of surface terrain, scientific understanding of how atmospheric and solar conditions affect the signal, and the precision of location and angle measurement. Each of these components is imperfect, and the degree of imperfection can be related to conditions on the ground like topographic roughness or prevailing weather. The CDL interprets this raw data using a decision tree, trained on Farm Services Agency (FSA) Common Land Unit data, a non-random sample generated from crop support program administrative data and therefore skewed toward subsidized crops (USDA NASS, 2024). For California as a whole in

[24]At the time of this writing, 2020-2023 are directly downloadable from their website, and 2018 and 2019 are available through archived copies via Wayback Machine. 2024 and 2025 are only available in provisional form.

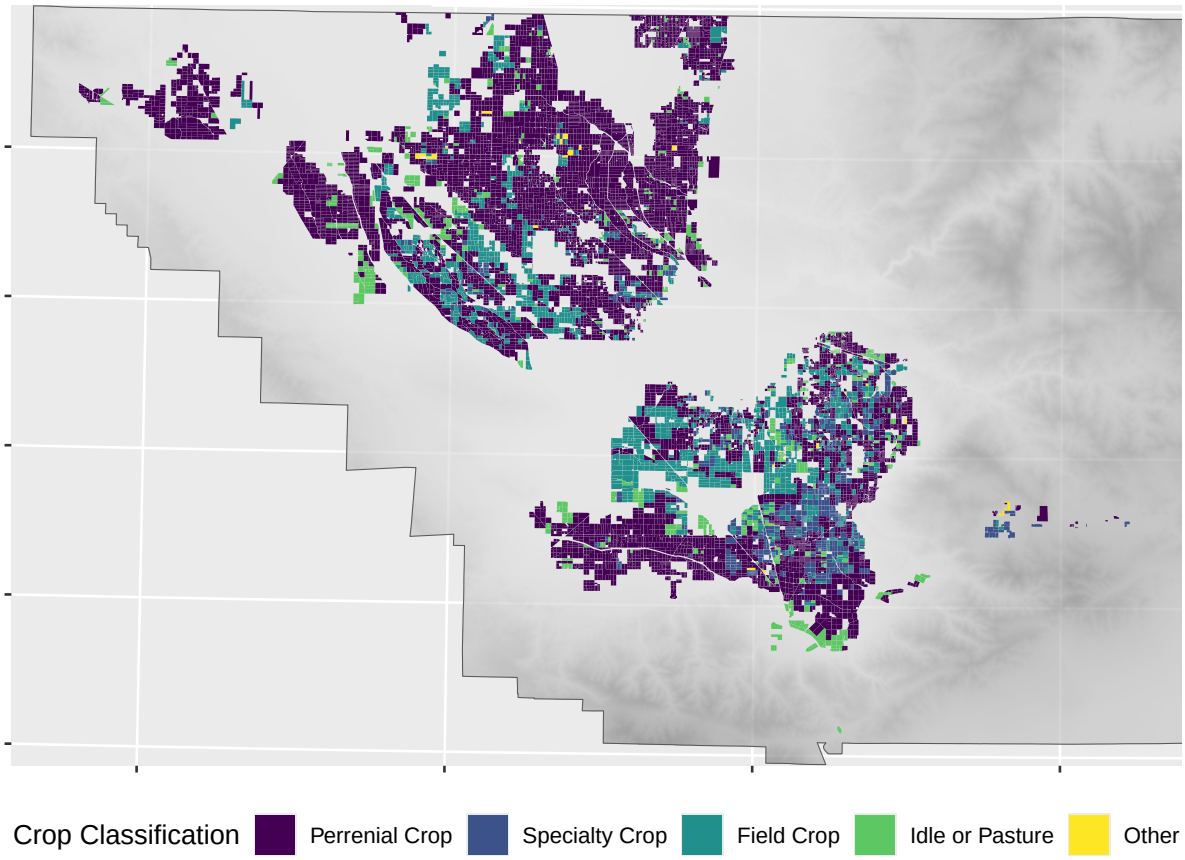


Figure 6: Map of High-Level Crop Categories on Kern County Fields, 2022

NOTES – Shapefiles for fields are the USDA Common Land Units. Crop categories are aggregated from granular crop data provided by the Kern County Department of Agriculture and Measurement Standards.

2023, the CDL reports an overall accuracy of 80.7% for FSA-covered crops, but this masks substantial heterogeneity: widespread crops like rice achieve 98.2% producer accuracy, while specialty crops such as cucumbers or watermelons can be below 40%. User accuracies follow a similar pattern.[25]

The CDL is a raster dataset, whereas the Kern County dataset is a shapefile, and the two use different crop categorizations. A typical approach for linking these data in multiple imputation would assign each shapefile polygon a crop based on the pixels it overlaps. In this application, however, I avoid directly privileging the Kern County shapefiles to prevent a scenario where the CDL appears inaccurate simply because it is evaluated against the Kern County data's own boundaries. Instead, before carrying out any multiple imputation, I link both datasets to the USDA Common Land Unit (CLU) shapefiles, which represent single, contiguously-managed fields—the natural unit for modeling crop choice. For each field, the CDL value is assiged as the plurality of underlying pixels, while the Kern County value is assigned as the plurality of acreage from intersecting field polygons. To reconcile differences in crop classification schemes, I aggregate crops to higher-level categories (e.g., broccoli and cabbage are grouped as "Cole Crops"). This aggregation will only make the CDL appear more accurate, as misclassifications between crops within the same category are treated as correct.

Field-level comparison between the CDL and the Kern County ground truth data reveals substantial misclassification. Figure 7 shows a confusion matrix for these two sets of data for 2022. Each cell is a count of fields, cells along the diagonal of this matrix correspond to fields correctly classified in the CDL, all other cells represent misclassifications. Percentages across the top show producer accuracies—the percentage of fields actually growing a crop category that are classified as growing that crop category in the satellite data.

[25]Producer accuracy is the percent of actual fields of a crop that are correctly classified. Even relatively common crops for California like onions can have strikingly low producer accuracies (63.9 percent). User accuracy is the percent of classified fields of a crop that are correctly classified.

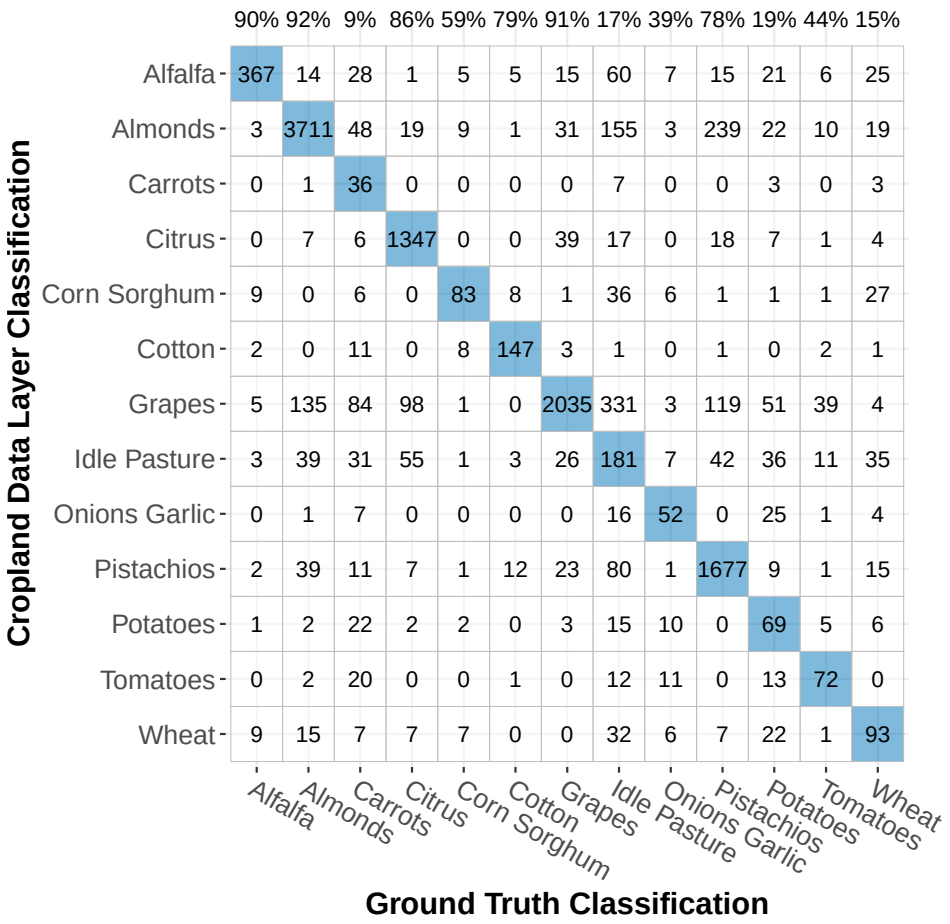


Figure 7: Confusion Matrix for Kern County Fields, 2022

NOTES – Data from the Kern County Department of Agriculture and Measurement Standards serves as ground truth. The ground truth data and the CDL raster data are intersected with USDA Common Land Unit Fields to produce these field counts. Individual crops are aggregated until categories perfectly match between the two datasets. Cell values correspond to counts of individual fields. Percentages across the top are producer accuracies. Crop categories making up less than one percent of Kern County field acreage are excluded; this causes minor disagreement between the producer accuracy percentages (which do include these categories) and values shown in the matrix. Correct classifications fall along the main diagonal, all other cells indicate misclassification.

Overall accuracy of the CDL in Kern County in 2022 is 71.6 percent by field count, rising only to 72.3 percent by acreage. Some crop categories never appear in the CDL despite being planted on dozens of fields (e.g. cole crops), others are routinely misclassified with producer accuracies well below 50 percent (e.g. carrots). Idle and pasture land is particularly difficult to classify correctly, a common feature in crop satellite data, since distinguishing between grassland pasture and grain crop production has more to do with management than attributes of the plants visible from satellite imagery. Many perennial crops are routinely misclassified as idle and vice versa, perhaps because immature orchards look more like bare fields than fully developed orchards to a machine classifier's top-down view.

Crop Choice Model

Using these data, I examine how year-to-year variation in water deliveries to Kern County water districts affects farmers' crop choices. In California, and particularly in Kern County, most agricultural water is delivered through water districts, which either hold water rights directly or maintain long-term contracts with the Central Valley Project and State Water Project. These two projects are extensive systems of dams, canals, and pumps that move water from the wetter north to the drier, intensively farmed south. Direct water rights provide stable access to pre-specified volumes of surface water, while project contracts provide access to a stable *share* of project water in each year. These institutions create both long-run differences in water availability across water districts and substantial year-to-year variation within districts. Long-run differences are clearly well-known to local farmers, and short-run variation, largely driven by variation in winter snow pack in the Sierra Nevada, is largely foreseeable by the time of planting.

Farms typically use water from their district allocation preferentially, with groundwater serving as a backstop. This water is cheap for farmers—water districts price water like

a cost-covering utility rather than charging for the water's real marginal cost.[26] During the study period, the main cost of groundwater extraction was electricity for pumping, a non-trivial cost given that Kern County's over-extracted groundwater is often hundreds of feet below the surface.[27]

A farmer deciding what to plant faces substantially different water availability constraints depending on what water district they live in and snow pack conditions over the preceding winter. The feasibility and profitability of various planting choices depend on the water requirements of that crop, how much of that requirement can be met with district water, and how costly additional backstop water will be. Short-run variability in district water deliveries impacts cropping decisions and is an important component of how drought impacts agriculture. Understanding how short-run water variability influences crop choice is crucial for evaluating drought impacts and informing adaptive water management strategies.

Given the diversity of crops grown in Kern County, a fully granular crop choice model is infeasible. I aggregate crops to four broad categories: perennial crops (e.g. almonds, oranges, wine grapes), specialty crops (e.g. carrots, lettuces, watermelons), field crops (e.g. grains, potatoes, dry beans), and fallow or pasture land. Parcels with non-agricultural or unconventional agricultural uses (e.g. developed land, nursery greenhouses) are excluded. This aggregation eliminates some misclassification error, but as the earlier confusion matrix showed, substantial measurement error remains in the CDL even at this higher level. Figure 8 shows the confusion matrix for these higher level aggregates.

I model the farmer's reaction to short-run water availability by specifying latent per-acre profit, where the crop $j \in \{Perennial, Specialty, Field, Fallow\}$ planted on field i in water district d in year t is chosen to maximize y_{ijt}:

$$y_{ijt} = \beta\ p_{jt} + \gamma_j \text{Water Delivered per Acre}_{dt} + \theta_{jd} + e_{ijt}$$

[26] The water districts themselves do not pay the marginal cost of water either, since they take their water from rights they own or contracts with the two projects.

[27] Groundwater pumping costs scale with the square of the depth to groundwater.

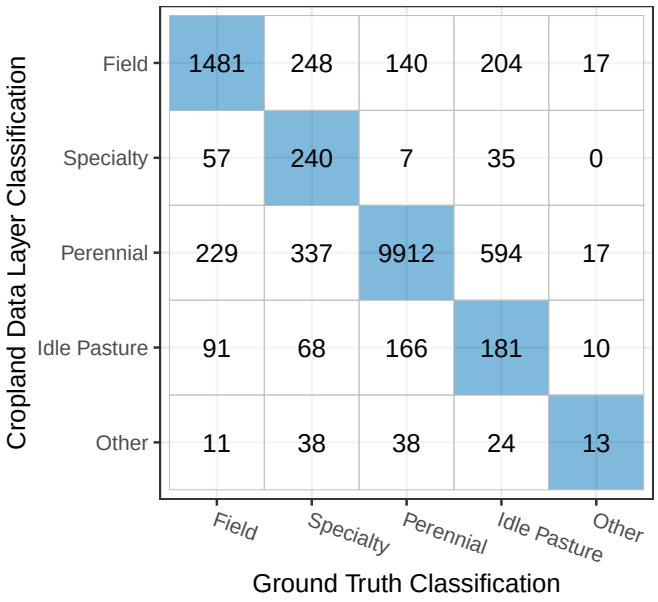


Figure 8: Confusion Matrix for Kern County Fields, 2022

NOTES – Data from the Kern County Department of Agriculture and Measurement Standards serves as ground truth. The ground truth data and the CDL raster data are intersected with USDA Common Land Unit Fields to produce these field counts. Cell values correspond to counts of individual fields. Correct classifications fall along the main diagonal, all other cells indicate misclassification.

where Fallow is the holdout category. Since reliable price series are unavailable for most specialty crops, I proxy p_{jt} using revenue per acre calculated from Kern County's annual crop reports, aggregated to the high-level crop categories with acreage weights. Water deliveries include water delivered to district d by the Central Valley Project and State Water Project. Collection of this data follows the procedures set forth by Hagerty (2024). District $\times$ crop-specific intercepts capture time invariant crop-specific costs in a way that allows differences across districts. This captures both long-run differences in district water availability and the costs of groundwater pumping, since depth to groundwater is highly spatially correlated and districts are, for the most part, geographically contiguous. My data cover cropping choices from 2018-2023 for 14917 fields across 14 water districts.[28] Though this is a relatively short panel, owing to the availability of the Kern County data, this timeframe does span both dry and wet years.

A farmer chooses crop j when $y_{ijt} > y_{ikt}$ for all other k, i.e.

$$\mathbb{P}(j = a) = \mathbb{P}(y_{iat} > y_{ibt} \ \forall \ b)$$

$$= \mathbb{P}(\beta(p_{at} - p_{bt}) + (\gamma_a - \gamma_b)(\text{Water Delivered per Acre}_{dt}) + (\theta_{ad} - \theta_{bd}) > e_b - e_a \ \forall \ b)$$

Assuming e_{ijt} is distributed Gumbel$(0, 1)$, this is a multinomial logit model, paralleling a random utility framework for a consumer's problem.

I estimate the crop choice model using three approaches:

1. Complete True Data: The model estimated using the full Kern County dataset.

2. Complete Satellite Data: The model estimated using exclusively CDL data.

3. Multiple Imputation: The model estimated using a random 10 percent sample of

[28]The included water districts are Arvin-Edison, Semitropic, Wheeler Ridge-Maricopa, Kern Delta, North Kern, Cawelo, Shafter-Wasco, Belridge, Rosedale-Rio Bravo, Buena Vista, Kern-Tulare, Berrenda Mesa, Delano-Earlimart, and Tehachapi-Cummings.

the Kern County data and the complete CDL data with multiple imputation. I set $\{k^{(\mathcal{X})}, k^{(\mathcal{Z})}\} = \{25, 10\}$ (calibrated to ensure truncation is not excessive). The model is iterated until convergence before generating 100 imputations. The imputation model includes indicators the crop data and indicators for district and year, omitting the other variables because they are deterministic functions of district and year. This is necessary because modelling the relationship of the district and year indicators non-parametrically already captures all variation at the district, year, and district $\times$ year level, as mentioned in the earlier section on panel data estimates. The water deliveries and revenue per acre variables are reintroduced when estimating the crop choice model.

Table 1 reports the estimated share of fields in each crop category in 2022 under each method. Both multiple imputation models reproduce crop shares that closely match the true data, outperforming the satellite only estimates.

Table 2 reports the estimated coefficients for key variables in the crop choice model. Relative to the model estimated with full true data, the CDL only estimates exhibit sign reversal and spurious significance. The results from multiple imputation, by contrast, align more closely with the true data. Estimates from the true data fall within the confidence interval for the multiple imputation estimates and vice versa. This is in contrast to the naive results, which exhibit spurious significance and in at least one case reject the point estimate from the true data model.

The empirical results highlight the potential for multiple imputation to serve as a bridge between high-quality microdata and large-scale predicted data. By leveraging flexible imputation modeling, the approach recovers not only accurate descriptive statistics but also valid structural parameters from highly flawed predicted satellite data.

Table 1: Percent of Kern County Fields in Each Crop Category (2022)

	True Data	Satellite Data	Multiple Imputation
	(1)	(2)	(3)
Perennial Crops	71.4%	78.3%	71.8%
	(0.4)	(0.4)	(0.4)
Specialty Crops	14.1%	15.7%	14.6%
	(0.3)	(0.3)	(0.3)
Field Crops	6.8%	2.5%	6.7%
	(0.2)	(0.1)	(0.2)
Idle or Pasture	7.6%	4.0%	6.9%
	(0.2)	(0.2)	(0.2)
Validation Share	1	0	0.1
Imputations	NA	NA	100

Table 2: Crop Choice Model Results for Key Variables

	True Data	Satellite Data	Multiple Imputation
	(1)	(2)	(3)
Revenue per Acre ($1000s)	0.007	-0.017^{***}	0.007
	(0.005)	(0.006)	(0.010)
Delivery per Acre (Field, af)	0.015	-0.089^{**}	0.016
	(0.033)	(0.036)	(0.068)
Delivery per Acre (Specialty, af)	0.067^{*}	0.084^{**}	0.024
	(0.035)	(0.042)	(0.073)
Delivery per Acre (Perennial, af)	0.010	−0.015	0.054
	(0.026)	(0.026)	(0.057)
Unique Field × Years	72535	72151	Varies
Validation Share	1	0	0.1
Imputations	NA	NA	100

Note: $p < 0.001^{***}$, $p < 0.05^{**}$, $p < 0.1^{*}$. Each estimate is from a multinomial logit crop choice model. Each model also contains a water district × crop-specific intercept.

Conclusion

Predicted categorical data, such as crop classification from satellite imagery or sentiment labels inferred from natural language processing of text, are now integral to empirical research across economics and the broader social sciences. Yet these data inevitably contain misclassification, and naively treating predictions as truth can induce substantial biases in downstream analyses. This paper develops a general, implementable method for correcting such biases using multiple imputation informed by a limited validation sample. The method requires only a small subset of correctly labeled observions and does not depend on knowing or specifying the form of the misclassification process.

By using a hierarchical Dirichlet process mixture model, a form of Bayesian nonparametric model, for imputation, the proposed approach flexibly captures the joint distribution of predicted and true categorical variables, as well as the dependence of this joint distribution on other covariates. In both simluations and an empirical application estimating the marginal revenue product of short-run water in Kern County agriculture, the multiple imputation method substantially reduces bias relative to analyses that rely on predicted data alone. Even when validation samples are limited or non-representative, the approach delivers meaningful improvements over naive estimation. These improvements are realized without requiring the researcher to retrain the prediction model or alter the code used for model estimation.

To take proper advantage of the research opportunities that predicted data make available, addressing misclassification must become standard practice in applied work. The approach presented here offers a practical and widely applicable solution: it treats predicted data as informative but uncertain, and integrates that uncertainty into inference through multiple imputation.

Appendix 1: Gibbs Sampler for the HCMM-LD

Before introducing the Gibbs Sampler, I first cover the remaining priors and hyperparameters that are not fully described in the main text.

As priors for the parameters in the X component, I use

$$\psi_s^{(j)} \overset{\text{iid}}{\sim} \text{Dirichlet}(\gamma_{s1}^{(j)}, \ldots, \gamma_{sd_j}^{(j)})$$

with each hyperparameter γ set equal to 1 to initialize. Posteriors are insensitive to this choice versus other minimally informative priors.

For the Y component, I use

$$\{(B_r, \Sigma_r)\} \overset{\text{iid}}{\sim} \text{MatN}(B_0, I, T_B) \times \text{IW}(q+1, \Sigma)$$

$$(B_0, \Sigma) \sim \text{MatN}(0, I, \sigma_{0\beta}^2 I) \times W\left(q+2, \frac{1}{q+1} I\right)$$

$$T_B = \text{diag}(1/\tau_1, \ldots, 1/\tau_q)$$

with priors for each $\tau_1, \ldots, \tau_q$ drawn independently from $G(0.5, 0.5)$ (with the second parameter refering to rate) and $\sigma_{0\beta}^2 = 10$. Each of these choices match those made by Murray and Reiter (2016) and are intended to be uninformative.

(In the code and package implementation, the Wishart hyperprior Σ is used only when the continuous data pass a degeneracy test. If any value of a continuous variable is repeated $q+2$ or more times, Σ is initialized and fixed at cI with $c = 0.25$, adjustable in the function call. This form of degeneracy can eliminate within-compnent variation and create an improper posterior. The fixed Σ bounds Σ_r below by $\Sigma/(n_r + q + 1)$.)

With these preliminaries, I now describe the Gibbs Sampler. Note that this only contains slight variations from Murray and Reiter (2016), mostly to simplify the process for the context

of estimating only on the complete cases.

1. Draw top-level component membership

For each observation, sample Z_i from

$$\mathbb{P}(Z_i = z | S_i = s, R_i = r) \propto \lambda_z \phi^{(\mathcal{X})}_{zs} \phi^{(\mathcal{Y})}_{zr}$$

2. Draw continuous and categorical component membership

For each observation sample s_i and r_i from

$$\mathbb{P}(S_i = s | Z_i = z, X_i = x_i, \cdot) \propto \phi^{(\mathcal{X})}_{zs} \prod_{j=1}^{p} \psi^{(j)}_{s x_{ij}}$$

$$\mathbb{P}(R_i = r | Z_i = z, Y_i = y_i, X_i = x_i, \phi^{(\mathcal{Y})}_z, B_h, \Sigma_h) \propto \phi^{(\mathcal{Y})}_{zr} N(y_i; D(x_i) B_r, \Sigma_r)$$

3. Update the within-component prior distributions and draw parameters

Draw component parameters for each $1 \leq s \leq k^{(\mathcal{X})}$ and $1 \leq j \leq p$ from

$$\psi^{(j)}_s \sim \text{Dirichlet}\left(\gamma_{s1} + \sum_{i=1}^{n} \mathbf{1}(S_i = s, X_{ij} = 1), \ldots, \gamma_{sd_j} + \sum_{i=1}^{n} \mathbf{1}(S_i = s, X_{ij} = d_j)\right)$$

and for each $1 \leq r \leq k^{(\mathcal{Y})}$ and $1 \leq v \leq q$ let B_{rv} be the vth column of B_r and sample

$$B_{rv} \sim N((\tau_v I + \mathbf{D}'_r \mathbf{D}_r / \tilde{\sigma}^2_{rv})^{-1} (\tau_v B_{0v} + \mathbf{D}'_r \tilde{\mathbf{y}}_{rv} / \tilde{\sigma}^2_{rv}), (\tau_v I + \mathbf{D}'_r \mathbf{D}_r / \tilde{\sigma}^2_{rv})^{-1})$$

where $\mathbf{D}_r$ is the matrix obtained by stacking the vector $\{D(x_i) : R_i = r\}$, $\tilde{\mathbf{y}}_{rv}$ is the vector obtained by concatenating $\{y_{iv} - \tilde{\mu}_{iv} : R_i = r\}$ and $\tilde{\mu}_{iv}$ and $\tilde{\sigma}^2_{rv}$ are parameters of the

conditional normal distribution for $(Y_{iv}|Y_{i/v}, R_i = r)$, where

$$(Y_{iv}|Y_{i/v}, R_i = r) \sim N(D(X_i)B_{rv} + \tilde{\mu}_{iv}, \tilde{\sigma}^2_{rv})$$

and finally for each $1 \leq r \leq k^{\mathcal{Y}}$ sample

$$\Sigma_r \sim IW(q + 1 + \sum_{i=1}^{n} \mathbf{1}(R_i = r), \Sigma + S_r)$$

where

$$S_r = \sum_{i:R_i=r} (Y_i - D(x_i)B_r)(Y_i - D(x_i)B_r)'$$

These parameter draws will be used in the next iteration of Step 2.

4. Update and draw hyperparameters

For each entry of B_0 sample

$$B_{0jv} \sim N\left((k^{(\mathcal{Y})}\tau_v + 1/\sigma^2_{0\beta})^{-1}\tau_v \sum_{r=1}^{k^{(\mathcal{Y})}} B_{rjv}, (k^{(\mathcal{Y})}\tau_v + 1/\sigma^2_{0\beta})^{-1}\right)$$

and for $1 \leq v \leq q$ sample

$$\tau_v \sim G\left(\frac{a_\tau + k^{(\mathcal{Y})}p^*}{2}, \frac{b_\tau + \sum_{r=1}^{k^{(\mathcal{Y})}}(B_{rv} - B_{0v})'(B_{rv} - B_{0v})}{2}\right)$$

If the degeneracy test mentioned in step 1 is passed, draw

$$\Sigma \mid \Sigma_r \sim W\left(q + 2 + K^{(\mathcal{Y})}(q + 1), \left[\left[\frac{1}{q+1}I\right]^{-1} + \sum_{r=1}^{k^{(\mathcal{Y})}} \Sigma_r^{-1}\right]^{-1}\right)$$

These parameter draws will be used in the next iteration of Step 3.

5. Draw stick-breaking parameters

Resample λ from

$$\xi_z^{(\mathcal{Z})} \sim \text{Beta}\left(1 + m_z, \alpha + n - \sum_{l=1}^{z} m_l\right)$$

for $1 \leq z \leq k^{(\mathcal{Z})} - 1$, with $\xi_{k^{(\mathcal{Z})}}^{(\mathcal{Z})} = 1$, and where $m_z = \sum_{i=1}^{n} \mathbf{1}(Z_i = z)$ and set $\lambda_z = \xi_z^{(\mathcal{Z})} \prod_{l<z} \left(1 - \xi_l^{(\mathcal{Z})}\right)$.

For $1 \leq z \leq k^{(\mathcal{Z})}$, iterate over $1 \leq s \leq k^{(\mathcal{X})} - 1$ and $1 \leq r \leq k^{(\mathcal{Y})} - 1$ sampling

$$\xi_{zs}^{(\mathcal{X})} \sim \text{Beta}\left(1 + t_{zs}^{(\mathcal{X})}, \beta^{(\mathcal{X})} + m_z - \sum_{l=1}^{s} t_{zl}^{(\mathcal{X})}\right)$$

$$\xi_{zr}^{(\mathcal{Y})} \sim \text{Beta}\left(1 + t_{zr}^{(\mathcal{Y})}, \beta^{(\mathcal{Y})} + m_z - \sum_{l=1}^{r} t_{zl}^{(\mathcal{Y})}\right)$$

where $t_{zs}^{(\mathcal{X})} = \sum_{i=1}^{n} \mathbf{1}(Z_i = z)\mathbf{1}(S_i = s)$ and $t_{zr}^{(\mathcal{Y})}$ is defined comparably. Generate $\phi_{zs}^{(\mathcal{X})}$ and $\phi_{zr}^{(\mathcal{Y})}$ from the ξs in the same was as in initialization.

These parameters will be used in the next iteration of Step 1.

6. Draw stick-breaking hyperparameters

Let $\mathcal{Z}_{occ}$ be the set of occupied top-level clusters with n_{occ} equal to the cardinality of $\mathcal{Z}_{occ}$. Sample

$$\alpha \sim G(a_0 + k^{(\mathcal{Z})} - 1, b_0 - \log(\lambda_{k^{(\mathcal{Z})}}))$$

$$\beta^{(\mathcal{X})} \sim G\left(a^{(\mathcal{X})} + n_{occ}(k^{(\mathcal{X})} - 1), b^{(\mathcal{X})} - \sum_{z \in \mathcal{Z}_{occ}} \log\left(\phi_{zk^{(\mathcal{X})}}^{(\mathcal{X})}\right)\right)$$

$$\beta^{(\mathcal{Y})} \sim G\left(a^{(\mathcal{Y})} + n_{occ}(k^{(\mathcal{Y})} - 1), b^{(\mathcal{Y})} - \sum_{z \in \mathcal{Z}_{occ}} \log\left(\phi_{zk^{(\mathcal{Y})}}^{(\mathcal{Y})}\right)\right)$$

I set $a^{(\mathcal{X})} = a^{(\mathcal{Y})} = 100$ and $b^{(\mathcal{X})} = b^{(\mathcal{Y})} = 50$, which keeps β away from values where new components are never born.

These hyperparameters will be used in the next iteration of Step 5.

To assess convergence, I compute the log-likelihood of the ground truth data at each every iteration of the model, then use this as the scalar estimand for computing a split Gelman-Rubin statistic (Gelman and Rubin, 1992; Gelman et al., 2014). While the typical advice is to compute Gelman-Rubin statistics for every parameter of interest in the model, in this use case the parameters of interest are the conditional probabilities of misclassification for every observation, much too large and unwieldy of a set of statistics to jointly consider. The log-likelihood of the data is a good estimand for assessing this statistic because computing it uses all of the parameters of interest in the model. Since the log-likelihood does not depend on actual component assignment, only probabilities of component assignments, it will not suffer from instability due to the label switching problem discussed by Jasra et al. (2005).

Appendix 2: Consistency of the HCMM-LD

This appendix establishes that the HCMM-LD, fit to validation data, consistently estimates the joint distribution of the data and (importantly for the application to multiple imputation) the conditional distribution of the true categorical variable given the observed variables. The conclusion here supports Assumption A3.4 in the proof of the consistency of the multiple imputation estimator in Appendix 3. In both cases, consistency is in the L^1 sense.

The argument combines existing results on the consistency of Bayesian non-parametric models for multivariate continuous and multivariate discrete data separately. The HCMM-LD's marginal distribution for the categorical variables is a mixture of product multinomial distributions, whose posterior consistency is proven by Dunson and Xing (2009). The conditional distribution for the continuous variables (given the categorical ones) is a mixture of normal distributions, whose posterior consistency is established by Wu and Ghosal (2010). I invoke their results as assumptions. The arguments below show that the consistency of these building blocks implies posterior consistency for the joint density and that the consistency of the joint density implies the weighted consistency of any conditional probability mass function derived from it. I then show when the Kullback-Leibler support condition underlying the building blocks holds. Finally, I connect these results to Appendix 3.

Setup and Notation

As in the main text, collect all categorical variables in $X = (Z, \tilde{Z}, X_1, \ldots, X_{p-2})$, coded as indicators so that $x \in \mathcal{X} = \{0,1\}^d$. Collect the continuous variables in Y, which takes values in $\mathbb{R}^q$. Let $f_0(x,y)$ be the true joint density with respect to the product of the counting measure on $\mathcal{X}$ and the Lebesgue measure on $\mathbb{R}^q$, let $h_0(x) = \int f_0(x,y)\, dy$ be the true marginal probability mass function of X, and let $f_0(y \mid x) = f_0(x,y)/h_0(x)$ be the true conditional density of Y given that $X = x$, so that $f_0(x,y) = f_0(y \mid x)h_0(x)$.

The validation data are n i.i.d. draws from f_0 (the final part of this section generalizes to

ignorable sampling designs). Let $f_n(x, y)$ be a draw from the HCMM-LD posterior given these data, with $h_n(x)$ its marginal probability mass function for X and $f_n(y \mid x)$ its conditional density for Y. The structure of the model gives $f_n(x, y) = f_n(y \mid x)h_n(x)$ with $h_n(x)$ a mixture of product multinomial distributions and $f_n(y \mid x)$ a mixture of normal distributions. Let Π denote the prior on joint densities induced by the model.

Assumptions

Assumption A2.1 (Consistency of the conditional density of Y)**.**

$$\sum_{x \in \mathcal{X}} h_0(x) \int |f_n(y \mid x) - f_0(y \mid x)| \ dy \xrightarrow{p} 0$$

Wu and Ghosal (2010) establish that $\int |f_n(y \mid x) - f_0(y \mid x)| \ dy \xrightarrow{p} 0$ for each fixed x. The weighted sum in Assumption A2.1 follows because $\mathcal{X}$ is finite and $h_0(x) \leq 1$. The finiteness of $\mathcal{X}$ is essential to this construction of the proof.

Assumption A2.2 (Consistency of the marginal probability mass function of X)**.**

$$\sum_{x \in \mathcal{X}} |h_n(x) - h_0(x)| \xrightarrow{p} 0$$

This is the result of Dunson and Xing (2009), which holds for every probability mass function h_0 on the finite set $\mathcal{X}$.

Assumption A2.3 (Kullback-Leibler support)**.** For every $\epsilon > 0$, $\Pi(f : D_{KL}(f_0||f) < \epsilon) > 0$.

Consistency of the Joint Density

Proposition A2.4 (Consistency of the joint density)**.** *Under Assumptions A2.1 and A2.2,* $||f_n(x, y) - f_o(x, y)||_1 \xrightarrow{p} 0$.

Proof. For every (x,y),

$$f_n(x,y) - f_0(x,y) = [f_n(y \mid x) - f_0(y \mid x)]h_n(x) + f_0(y \mid x)[h_n(x) - h_0(x)]$$

Taking absolute values, summing over x, integrating over y, and using $\int f_0(y \mid x)\, dy = 1$,

$$||f_n(x,y) - f_o(x,y)||_1 \leq \sum_{x \in \mathcal{X}} h_n(x) \int |f_n(y \mid x) - f_0(y \mid x)|\, dy + ||h_n(x) - h_0(x)||_1$$

The second term $\xrightarrow{p} 0$ by Assumption A2.2. Recognizing that $h_n(x) = h_0(x) + [h_n(x) - h_0(x)]$ and that $\int |f_n(y \mid x) - f_0(y \mid x)|\, dy \leq 2$,

$$\sum_{x \in \mathcal{X}} h_n(x) \int |f_n(y \mid x) - f_0(y \mid x)|\, dy \leq \sum_{x \in \mathcal{X}} h_0(x) \int |f_n(y \mid x) - f_0(y \mid x)|\, dy + 2||h_n(x) - h_0(x)||_1$$

which $\xrightarrow{p} 0$ by Assumptions A2.1 and A2.2. □

Consistency of the Conditional Probability Mass Function

Partition $x = (x_a, x_b)$ where x_a collects the indicators of the categorical variable being imputed (Z in the main text) and x_b are the remaining categorical indicators, with $\mathcal{X}_a$ and $\mathcal{X}_b$ their corresponding sets of values. Write $v = (x_b, y)$ for the conditioning variables, which take values in $\mathbb{V} = \mathcal{X}_b \times \mathbb{R}^q$. The imputations are draws from the posterior conditional of X_a given V. Define the true density of V and the true conditional probability mass functionf of X_a given V by

$$s_0(v) = \sum_{x_a \in \mathbb{X}_a} f_0(x_a, x_b, y), \quad g_0(x_a \mid v) = \frac{f_0(x_a, x_b, y)}{s_0(v)}$$

with $g(x_a \mid v$ defined arbirarily where $s_0(v) = 0$. Define s_n and g_n in the same way from f_n. For draws from the HCMM-LD, $s_n(v) > 0$ everywhere, so g_n is well-defined.

Proposition A2.5 (Consistency of the weighted conditional probability mass function)**.**

Under Assumptions A2.1 and A2.2,

$$\sum_{x_b \in \mathcal{X}_b} \int \sum_{x_a \in \mathcal{X}_a} |g_n(x_a \mid v) - g_0(x_a \mid v)| s_0(v)\, dy \xrightarrow{p} 0$$

Proof. Since $g_n s_n = f_n$ and $g_0 s_0 = f_0$, for every (x_a, v),

$$[g_n(x_a \mid v) - g_0(x_a \mid v)] s_0(v) = [f_n(x_a, x_b, y) - f_0(x_a, x_b, y)] - g_n(x_a \mid v)[s_n(v) - s_0(v)]$$

Taking absolute values, summing over x_a, and using $\sum_{x_a \in \mathcal{X}_a} g_n(x_a \mid v) = 1$,

$$\sum_{x_a \in \mathcal{X}_a} |g_n(x_a \mid v) - g_0(x_a \mid v)| s_0(v) \leq \sum_{x_a \in \mathcal{X}_a} |f_n(x_a, x_b, y) - f_0(x_a, x_b, y)| + |s_n(v) - s_0(v)|$$

Summing over x_b and integrating over y, the first term is $||f_n - f_0||_1$. For the second, $|s_n(v) - s_0(v)| = |\sum_{x_a \in \mathcal{X}_a} f_n - f_0| \leq \sum_{x_a \in \mathcal{X}_a} |f_n - f_0|$, so after the summing and integration the second is also at most $||f_n - f_0||_1$. The right is then collectively at most $2||f_n - f_0||_1$, which $\xrightarrow{p} 0$ by Proposition A2.4. □

Weighting by s_0 is all that Appendix 3 requires: the imputation model only needs to be accurage on average over the conditioning variables under their true distribution.

Kullback-Leibler Support of the Prior

Both Assumptions A2.1 and A2.2 require that the model's prior put positive mass on every Kullback-Leibler neighborhood of the true distribution (Assumption A2.3). An exhaustive description of all densities f_0 for which the HCMM-LD prior satisfies this condition is beyond the scope of this paper, but existing results show that this includes a wide class. The Kullback-Leibler divergence from f_0 to a density $f(x, y) = f(y \mid x)h(x)$ can be decomposed:

$$D_{KL}(f_0||f) = \sum_{x \in \mathcal{X}} h_0(x) \log \frac{h_0(x)}{h(x)} + \sum_{x \in \mathcal{X}} h_0(x) \int f_0(y \mid x) \log \frac{f_0(y \mid x)}{f(y \mid x)}\, dy$$

Dunson and Xing (2009) show that a Dirichlet process mixture of product multinomials places positive mass on every Kullback-Leibler neighborhood of every probability mass function on $\mathcal{X}$. Mixtures of normal distributions are universal approximators of smooth densities, and several papers establish Kullback-Leibler support iof Dirichlet process mixtures of normals at densities satisfying mild regularity conditions, typically continuity and adequately light tails. Taken together, for every f_0 whose marginal probability mass function for the categorical variables and conditional density for the continuous variables satisfy such conditions, there are densities in the support for Π for which each term of the decomposition is below $\epsilon/2$, so Assumption A2.3 holds. One set of sufficient conditions is the following:

1. Positivity of the discrete support: $h_0(x) > 0$ for all $x \in \mathcal{X}$
2. Boundedness: $f_0(y \mid x)$ is bounded above for every x
3. Finite entropy: $\int f_0(y \mid x) |\log f_0(y \mid x)| \, dy < \infty$ for every x
4. Mild local variation: For some $\delta > 0$ and every x,

$$\int f_0(y \mid x) \log \frac{f_0(y \mid x)}{\phi_\delta(y \mid x)} \, dy < \infty, \quad \phi_\delta(y \mid x) = \inf_{||t-y||<\delta} f_0(t \mid x)$$

5. Tails: For some $\rho > 0$ and every x, $\int f_0(y \mid x) ||y||^{2(1+\rho)q} \, dy < \infty$

Condition (1) is not necessary for the consistency of h_n and is not imposed by Dunson and Xing (2009). Nor is it substantive, when some combinations of the categorical variables are impossible, $\mathcal{X}$ can be redefined as the support of h_0. Conditions (2)-(5) are the conditins of Wu and Ghosal (2010). They are mild, and densities that violate them are difficult to approximate by any method. Kleijn and van der Vaart (2006) show, moreover, that when f_0 lies outside the support of Π, the posterior concentrates on the density f^* in the support that minimizes $D_{KL}(f_0||f^*)$, which offers a certain sense of robustness even in the event that f_0 fails these mild conditions.

Relationship to Appendix 3

In the notation of Appendix 3, X_a is Z, the conditioning variables $V = (\tilde{Z}, X, Y)$ are $v = (x_b, y)$, the true conditional $g_0(x_a \mid v)$ is $g_0(z \mid v)$, and summing over x_b and integrating against $s_0(v)\ dy$ is integrating against $P_{0,V}$. The posterior draw g_n is the imputation conditional $\hat{g}^m$. Proposition A2.5 satisfies Assumption A3.4. To extend this appendix's proof to an ignorable sampling setting (following Assumption 2 in the main text), simply apply this appendix to $f_0^*(x, y) = \pi(v) f_0(x, y) / \mathbb{E}\pi(v)$ instead of f_0. This changes nothing in the conditional ($g_0^*(x_a \mid v) = g_0(x_a \mid v)$).

Appendix 3: Consistency of the Multiple Imputation Estimator

This appendix demonstrates the consistency of the multiple imputation estimator, in the sense that the estimator converges in probability to the same value as the infeasible estimator applied to data sampled from the true underlying population distribution. The consistency proof applies to a wide class of estimators of interest, including M-estimators (maximum likelihood, least squares, quantile regression), generalized method of moments estimators, continuous functions of sample means (ordinary or two-stage least squares), and plug-in estimators that are continuous functionals of the data distribution (quantiles). The proof does not attempt to exhaust all possible applications, and the multiple imputation estimator may still be consistent for estimators not captured by the proof's conditions.

The primary wrinkle in applying standard consistency proofs to the multiple imputation estimator is that the imputed data may contain combinations of values that have probability zero in the true distribution—this is because the imputed values are drawn from an estimated conditional distribution that only approaches the truth as the validation sample size grows. Therefore, whereas standard consistency proofs need only to control the behavior of the estimator over the support of the data's true distribution, these proofs must also control that behavior over imputed samples where impossible data combinations may linger in the support of the conditional distribution estimated by the imputation model.

The proof relies on a modified uniform law of large numbers, applied to imputed samples. It exploits the fact that only the finitely-valued variable Z is imputed, while all other variables are drawn from their true distributions, and requires only that the imputation model be accurate on average over the values it imputes. To assist with readability, the proofs first proceed under the assumption that the validation data are a simple random sample of the overall data (Assumption A3.2), but Corollary A3.18 shows that each of the consistency

proofs remains valid under ignorable validation sampling.

Setup and Notation

Z takes values in the finite set $\mathcal{Z}$ and the variables observed for every observation are collected in $V = (\tilde{Z}, X, Y)$ which takes values in a space $\mathbb{V}$. P_0 is the joint distribution of (Z, V) in the population, $P_{0,V}$ is the marginal distribution of V, and

$$g_0(z \mid v) = P_0(Z = z \mid V = v), \quad z \in \mathcal{Z}$$

is the true conditional probability mass function of Z. The data contain N units (Z_i, V_i, R_i) where $R_i = 1$ if unit i is in the validation sample (i.e. the true Z_i is observed). There are $N^* = \sum_i R_i$ units in the validation sample.

For imputations $m = 1, \ldots, M$, $\hat{g}^m(z \mid v)$ is the conditional probability mass function resulting from the m-th draw of the imputation model. The m-th completed dataset is $\{(Z_i^m, V_i)\}_{i=1}^N$, where $Z_i^m = Z_i$ when $R_i = 1$ and is drawn from $\hat{g}^m(\cdot \mid V_i)$ otherwise.

There are two sources of auxiliary randomness, $\mathcal{U}_S$ and $\mathcal{U}_I$, which are mutually independent with each other and with $(Z_{1:N}, V_{1:N}, R_{1:N})$. $\mathcal{U}_S$ is the randomness embedded in the sampler: draws for component memberships, stick-breaking weights, and component parameters. $\mathcal{U}_I = \{U_i^m\}$ is a collection of i.i.d. standard uniform variables used to draw Z_i^m from $\hat{g}^m(\cdot|V_i)$ for observations with $R_i = 0$.

The L_1 discrepancy of a draw of the imputation model at v is

$$d^m(v) = \sum_{z \in \mathcal{Z}} \left| \hat{g}^m(z \mid v) - g_0(z \mid v) \right|$$

where $0 \leq d^m(v) \leq 2$ due to the triangle inequality and $\mathcal{Z}$ being finite.

For the proofs, I will often work with functions of the completed data $h : \mathcal{Z} \times \mathbb{V} \times \mathcal{B} \to \mathbb{R}^k$,

where $\mathcal{B}$ is the parameter space. Combined with a probability mass function g on $\mathcal{Z}$, define

$$\bar{h}_g(v, \beta) = \sum_{z \in \mathcal{Z}} h(z, v, \beta) g(z \mid v)$$

and the envelopes

$$H_\beta(v) = \max_{z \in \mathcal{Z}} ||h(z, v, \beta)||, \quad H(v) = \max_{z \in \mathcal{Z}} \sup_{\beta \in \mathcal{B}} ||h(z, v, \beta)||$$

In the interest of simplicity, I first describe the proof assuming that the validation sample is a simple random sample of the overall data. A later extension will extend the proof to the case of validation data whose sampling probability depends on observed variables. The assumptions invoked here are:

Assumption A3.1 (Sampling)**.** (Z_i, V_i) for $i = 1, \ldots, N$ are i.i.d. from distribution P_0, which does not depend on N.

Assumption A3.2 (Simple random validation sample)**.** The validation indicators $R_{1:N}$ are independent of $\{(Z_i, V_i)\}$ and $N^* \xrightarrow{p} \infty$.

Assumption A3.3 (Imputation model fit with validation data)**.** Let $D_N^* = (R_i, R_i Z_i, R_i V_i)_{i=1,\ldots,N}$ be the validation data, which resides in a Borel subset $\mathbb{D}_N$ of Euclidean space.[29] Each draw of the imputation model has the form $\hat{g}^m(z \mid v) = G_N^m(z, v; D_N^*, \mathcal{U}_S)$, where G_N^m is jointly measurable with respect to the Borel σ-algebras on $\mathcal{Z} \times \mathbb{V} \times \mathbb{D}_N$ and the product σ-algebra on $[0, 1]^\infty$, the space of sequences in $[0, 1]$ from which $\mathcal{U}_S$ takes values.

Assumption A3.4 (Consistency of the imputation model)**.** For each m,

$$\Delta_N^m := \int d^m(v) \; dP_{0,V}(v) \xrightarrow{p} 0$$

[29] Multiplying observables by R_i masks values from non-validated observations. This implementation keeps D_N^* a fixed dimension even if N^* is random.

Assumptions A3.1 and A3.2 are the simple random sample counterparts of the sampling assumptions in the main text. These will be relaxed to accommodate ignorable sampling designs at the end of this appendix. Assumption A3.3 establishes general conditions on the form of the imputation model and clarifies that it is fit only with validation data, which simplifies the proof. Assumption A3.4 is proven for the HCMM-LD used in the main text in Appendix 2, but is kept general, allowing this proof to apply for any multiple imputation procedure with an imputation model that converges to the true data distribution in L^1 distance.

For the following three lemmas, let Assumptions A3.1–A3.4 hold and let $\mathcal{F}_N = \sigma(D_N^*, \mathcal{U}_S, V_{1:N})$.

Lemma A3.5 (Independence of validated values)**.** *Conditional on* $(V_{1:N}, R_{1:N})$*, the variables* $Z_{1:N}$ *are independent with* $\mathbb{P}(Z_i = z \mid V_{1:N}, R_{1:N}) = g_0(z \mid V_i)$.

Proof. Assumption A3.2 implies that conditioning on $R_{1:N}$ makes no difference, while Assumption A3.1 implies that V_i is the only relevant part of conditioning on $V_{1:N}$. □

Lemma A3.6 (Independence of imputed values)**.** *Each imputation conditional* $\hat{g}^m$ *is* $\mathcal{F}_N$*-measurable and the imputed values* $\{Z_i^m : R_i = 0\}$ *are independent conditional on* $\mathcal{F}_N$*, with* $\mathbb{P}(Z_i^m = z \mid \mathcal{F}_N) = \hat{g}^m(z \mid V_i)$.

Proof. The measurability of $\hat{g}^m$ is immediate from A3.3. The U_i^m determining draws from $\hat{g}^m(z \mid V_i)$ for each imputation are independent of $\mathcal{F}_N$, so the draws are conditionally independent with the stated probability. □

Lemma A3.7 (Consistency on imputed units)**.** *For each* m*,*

$$\hat{\Delta}_N^m := \frac{1}{N}\sum_{i=1}^{N}(1 - R_i)d^m(V_i) \xrightarrow{p} 0$$

Proof. Since $R_{1:N}$ is independent of the data and the units are i.i.d., the unvalidated units' V_i is independent of the validated units' data and of $\mathcal{U}_S$. D_N^* reveals only that $R_i = 0$ for such a unit. Therefore, the unvalidated V_i are i.i.d. with $V_i \mid D_N^*, \mathcal{U}_S, R_i = 0 \sim P_{0,V}$.

By Assumption A3.3, $d^m(v) = \sum_z |G_N^m(z, v; D_N^*, \mathcal{U}_S) - g_0(g \mid v)|$ is a Borel function of $(v, D_N^*, \mathcal{U}_S)$, so Fubini's theorem gives $\mathbb{E}[d^m(V_i) \mid D_N^*, \mathcal{U}_S] = \Delta_N^m$, where Δ_N^m is a function of $(D_N^*, \mathcal{U}_S)$. Since D_N^* fully reveals R_i,

$$\mathbb{E}[\hat{\Delta}_N^m \mid D_N^*, \mathcal{U}_S] = \frac{N - N^*}{N} \Delta_N^m \leq \Delta_N^m$$

Taking expectations, $\mathbb{E}[\hat{\Delta}_N^m] \leq \mathbb{E}[\Delta_N^m]$. $\mathbb{E}[\Delta_N^m] \to 0$ by Assumption A3.4 and bounded convergence (Δ_N^m integrates a function bounded above by 2 over a probability distribution). Then Markov's inequality gives $\mathbb{P}(\hat{\Delta}_N^m > \eta) \leq \mathbb{E}[\Delta_N^m]/\eta \to 0$ for every $\eta > 0$.

□

A Uniform Law of Large Numbers for Imputed Samples

Lemma A3.8 (Conditionally centered sums)**.** *Let Assumption A3.1 hold and let $\mathcal{G}_N$ be a σ-algebra for which $V_1, \ldots, V_N$, indicators $a_1, \ldots, a_N \in \{0, 1\}$, and the probabilities $G_i(z \mid V_i)$ (coming from probability mass functions $G_1, \ldots, G_N$ on $\mathcal{Z}$) are $\mathcal{G}_N$-measurable. Conditional on $\mathcal{G}_N$, let $\{\xi_i : a_i = 1\}$ be independent $\mathcal{Z}$-valued random variables with $\mathbb{P}(\xi_i = z \mid \mathcal{G}_N) = G_i(z \mid V_i)$ for each i with $a_i = 1$. Define the function*

$$A_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} a_i \left[h(\xi_i, V_i, \beta) - \bar{h}_{G_i}(V_i, \beta) \right]$$

For i with $a_i = 0$, ξ_i is arbitrary because they do not enter the expression.

(i) If $\mathbb{E}_{P_0} H_\beta(V) < \infty$, then $A_N(\beta) \xrightarrow{p} 0$.

(ii) If $\mathcal{B}$ is compact, $h(z, v, \beta)$ is continuous in β for every z and for $P_{0,V}$-almost every v,

and $\mathbb{E}_{P_0} H(V) < \infty$*, then* $\sup_{\beta \in \mathcal{B}} ||A_N(\beta)|| \xrightarrow{p} 0$.

Proof. (i) Fix β. Note that the expectation of each summand of $A_N(\beta)$ conditional on $\mathcal{G}_N$ is zero. For a fixed $K > 0$, split $h = h^{\leq K} + h^{>K}$, where $h^{\leq K}(z, v, \beta) = h(z, v, \beta)\mathbf{1}\{H_\beta(v) \leq K\}$. The splitting depends on v only, which is $\mathcal{G}_N$-measurable, so is deterministic given $\mathcal{G}_N$. Therefore the summands of both $A_N^{\leq K}(\beta)$ and $A_N^{>K}(\beta)$ remain conditionally centered and $A_N(\beta) = A_N^{\leq K}(\beta) + A_N^{>K}(\beta)$.

Conditional on $\mathcal{G}_N$, the summands of $A_N^{\leq K}(\beta)$ are independent and have a second moment of at most K^2, so $\mathbb{E}||A_N^{\leq K}(\beta)||^2 \leq K^2/N$.

For the other part, $||A_N^{>K}(\beta)|| \leq \frac{2}{N} \sum_i H_\beta(V_i)\mathbf{1}\{H_\beta(V_i) > K\}$, whose expectation is $2\mathbb{E}_{P_0}[H_\beta(V_i)\mathbf{1}\{H_\beta(V_i) > K\}]$.

So, by Chebyshev's, Markov's, and the triangle inequalities,

$$\mathbb{P}(||A_N(\beta)|| > 2\eta) \leq \mathbb{P}(||A_N^{\leq K}(\beta)|| > \eta) + \mathbb{P}(||A_N^{>K}(\beta)|| > \eta) \leq \frac{K^2}{N\eta^2} + \frac{2}{\eta}\mathbb{E}_{P_0}[H_\beta(V_i)\mathbf{1}\{H_\beta(V_i) > K\}]$$

The first term $\to 0$ as $N \to \infty$. The second term vanishes as $K \to \infty$ by dominated convergence.

(ii) For $\beta \in \mathcal{B}$ and $\delta > 0$, define

$$D_\delta(v, \beta) = \max_{z \in \mathcal{Z}} \sup_{\beta' \in \mathcal{B}: ||\beta' - \beta|| < \delta} ||h(z, v, \beta') - h(z, v, \beta)||$$

Then $D_\delta \leq 2H(V)$ and $\lim_{\delta \to 0^+} D_\delta(v, \beta) = 0$ for almost every v due to h's continuity and the finiteness of $\mathcal{Z}$. By dominated convergence, $\lim_{\delta \to 0^+} \mathbb{E}_{P_0} D_\delta(V, \beta) = 0$. Fix $\epsilon > 0$ and choose, for each β, a radius $\delta(\beta)$ with $\mathbb{E}_{P_0} D_{\delta(\beta)}(V, \beta) < \epsilon$. By compactness, finitely many of these open balls, centered at $\beta_1, \ldots, \beta_L \in \mathcal{B}$ cover $\mathcal{B}$. If β is in the l-th ball, then both $||h(\xi_i, V_i, \beta) - h(\xi_i, V_i, \beta_l||$ and $||\bar{h}_{G_i}(V_i, \beta) - \bar{h}_{G_i}(V_i, \beta_l)||$ are at most $D_{\delta_l}(V_i, \beta_l)$, no matter what value ξ_i takes or which probability mass function G_i is used. For any β,

$||A_N(\beta)|| \leq ||A_N(\beta_l)|| + ||A_N(\beta) - A_N(\beta_l)||$ by the triangle inequality. Therefore,

$$\sup_{\beta \in \mathcal{B}} ||A_N(\beta)|| \leq \max_l ||A_N(\beta_l)|| + 2 \max_l \frac{1}{N} \sum_{i=1}^{N} D_{\delta_l}(V_i, \beta_l)$$

The first term is $o_p(1)$ by part (i). The second converges in probability to $2 \max_l \mathbb{E}_{P_0} D_{\delta_l}(V, \beta_l) < 2\epsilon$, so $\mathbb{P}(\sup_{\beta \in \mathcal{B}} ||A_N(\beta)|| > 3\epsilon) \to 0$

□

Lemma A3.9 (Imputation model error)**.** *Let Assumptions A3.1-A3.4 hold. Define*

$$B_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} (1 - R_i)[\bar{h}_{\hat{g}^m}(V_i, \beta) - \bar{h}_{g_0}(V_i, \beta)]$$

(i) For any $\beta \in \mathcal{B}$ such that $\mathbb{E}_{P_0} H_\beta(V) < \infty$, $||B_N(\beta)|| \xrightarrow{p} 0$.

(ii) If $\mathbb{E}_{P_0} H(V) < \infty$, then $\sup_{\beta \in \mathcal{B}} ||B_N(\beta)|| \xrightarrow{p} 0$.

Proof. For every β, $||\bar{h}_{\hat{g}^m}(v, \beta) - \bar{h}_{g_0}(v, \beta)|| \leq \sum_z ||h(z, v, \beta)|| \; |\hat{g}^m(z \mid v) - g_0(z \mid v)| \leq H(v) d^m(v)$. Since $d^m(v) \leq 2$, for any $K > 0$, $H(v) d^m(v) \leq 2H(v) \mathbf{1}\{H(v) > K\} + K d^m(v)$, so

$$\sup_{\beta \in \mathcal{B}} ||B_N(\beta)|| \leq \frac{2}{N} \sum_{i=1}^{N} H(V_i) \mathbf{1}\{H(V_i) > K\} + K \hat{\Delta}_N^m$$

The first term $\xrightarrow{p} 2\mathbb{E}_{P_0} H(V) \mathbf{1}\{H(V) > K\}$ which is arbitrarily small for large enough K. The second term is $o_p(1)$ for any fixed K by Lemma A3.7. This proves (ii). Statement (i) follows the same argument, removing the supremum and replacing $H(V)$ with $H_\beta(V)$.

□

Proposition A3.10 (Uniform law of large numbers for imputed samples)**.** *Let Assumptions A3.1-A3.4 hold.*

(i) If $\mathbb{E}_{P_0} H_\beta(V) < \infty$, then $\frac{1}{N} \sum_{i=1}^{N} h(Z_i^m, V_i, \beta) \xrightarrow{p} \mathbb{E}_{P_0} h(Z, V, \beta)$

(ii) If $\mathcal{B}$ is compact, $h(z, v, \beta)$ is continuous in β for every z and for $P_{0,V}$-almost every v, and $\mathbb{E}_{P_0} H(V) < \infty$, then $\mathbb{E}_{P_0} h(Z, V, \beta)$ is continuous in β and

$$\sup_{\beta \in \mathcal{B}} \left\| \frac{1}{N} \sum_{i=1}^{N} h(Z_i^m, V_i, \beta) - \mathbb{E}_{P_0} h(Z, V, \beta) \right\| \xrightarrow{p} 0$$

Proof. Begin with the following decomposition, adding and subtracting $\bar{h}_{g_0}(V_i, \beta)$ for every unit and $\bar{h}_{\hat{g}^m}(V_i, \beta)$ for every imputed unit:

$$\frac{1}{N} \sum_{i=1}^{N} h(Z_i^m, V_i, \beta) = C_N(\beta) + \Xi_N(\beta) + A_N(\beta) + B_N(\beta)$$

where

$$C_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} \bar{h}_{g_0}(V_i, \beta) \qquad \text{(Observed data term)}$$

$$\Xi_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} R_i [h(Z_i, V_i, \beta) - \bar{h}_{g_0}(V_i, \beta)] \qquad \text{(Sampling noise in validated values)}$$

$$A_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} (1 - R_i) [h(Z_i^m, V_i, \beta) - \bar{h}_{\hat{g}^m}(V_i, \beta)] \qquad \text{(Simulation noise in imputed values)}$$

$$B_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} (1 - R_i) [\bar{h}_{\hat{g}^m}(V_i, \beta) - \bar{h}_{g_0}(V_i, \beta)] \qquad \text{(Imputation model error)}$$

Observed data term. $C_N(\beta)$ is an average of i.i.d. functions of V_i, dominated by $H_\beta(V)$ (or by $H(V)$), and $\mathbb{E}_{P_0} \bar{h}_{g_0}(V, \beta) = \mathbb{E}_{P_0} h(Z, V, \beta)$ by the law of iterated expectations. Under the assumption of (i), the weak law of large numbers applies. Under the assumptions of (ii), the summands of $C_N(\beta)$ are continuous (finite sums of continuous functions), so the uniform law of large numbers applies.

Sampling noise in validated values. This part $\to 0$ by Lemma A3.8 with $\mathcal{G}_N = \sigma(V_{1:N}, R_{1:N})$,

$\xi_i = Z_i$, $G_i = g_0$, and $a_i = R_i$. The conditional structure required is satisfied by Lemma A3.5.

Simulation noise in imputed values. This part $\to 0$ by Lemma A3.8 with $\mathcal{G}_N = \mathcal{F}_N$, $\xi_i = Z_i^m$, $G_i = \hat{g}^m$, and $a_i = 1 - R_i$.The conditional structure required is satisfied by Lemma A3.6.

Imputation model error. This part $\to 0$ by Lemma A3.9.

□

Consistency Proofs

For each of the following consistency theorems, let $\hat{\beta}^m$ denote the estimate obtained from the m-th imputed dataset and let $\hat{\beta}_{MI} = \frac{1}{M}\sum_{m=1}^{M} \hat{\beta}^m$.

Theorem A3.11 (M-estimators)**.** *Let* $\hat{Q}_N^m(\beta) = \frac{1}{N}\sum_{i=1}^{N} q(Z_i^m, V_i, \beta)$ *and* $Q_0(\beta) = \mathbb{E}_{P_0} q(Z, V, \beta)$*. Let Assumptions A3.1-A3.4 hold, and suppose* Q_0 *is uniquely minimized over* $\mathcal{B}$ *at* β_0*. Further suppose that one of the following holds:*

(i) $\mathcal{B}$ *is compact,* $q(z, v, \beta)$ *is continuous in* β *for every* z *and almost every* v*,* $\mathbb{E}_{P_0} \max_z \sup_{\beta \in \mathcal{B}} |q(z, V, \beta)| < \infty$*, and* $\hat{Q}_N^m(\hat{\beta}^m) \leq \inf_\beta \hat{Q}_N^m(\beta) + o_p(1)$*.*

(ii) $\mathcal{B}$ *is convex and* β_0 *is in its interior,* $q(z, v, \beta)$ *is convex in* β *for every* (z, v)*,* $\mathbb{E}_{P_0} \max_z |q(z, V, \beta)| < \infty$ *for all* $\beta \in \mathcal{B}$*, and* $\hat{\beta}^m$ *minimizes* $\hat{Q}_N^m$ *whenever a minimizer exists.*

Then $\hat{\beta}^m \xrightarrow{p} \beta_0$ *for each* m *and* $\hat{\beta}_{MI} \xrightarrow{p} \beta_0$ *for any* M*.*

Proof. Under (i), Proposition A3.10 with $h = q$ gives $\sup_{\beta \in \mathcal{B}} ||\hat{Q}_N^m(\beta) - Q_0(\beta)|| \xrightarrow{p} 0$. @Newey94 Theorem 2.1 applies, so $\hat{\beta}^m \xrightarrow{p} \beta_0$.

Under (ii), Proposition A3.10 with $h = q$ gives $\hat{Q}_N^m(\beta) \xrightarrow{p} Q_0(\beta)$. @Newey94 Theorem 2.7 applies, so a minimizer $\hat{\beta}^m$ exists with probability approaching one and $\hat{\beta}^m \xrightarrow{p} \beta_0$.

The average of a finite sum of values each converging in probability to β_0 also converges in probability to β_0, so $\hat{\beta}_{MI} \xrightarrow{p} \beta_0$. □

Theorem A3.12 (GMM). *Let* $\hat{\varphi}_N^m(\beta) = \frac{1}{N}\sum_{i=1}^N \varphi(Z_i^m, V_i, \beta)$ *for a moment function* φ *and let* $\hat{\beta}^m$ *satisfy* $\hat{Q}_N^m(\hat{\beta}^m) \leq \inf_\beta \hat{Q}_N^m(\beta) + o_p(1)$ *where* $\hat{Q}_N^m(\beta) = \hat{\varphi}_N^m(\beta)'\hat{\Omega}^m\hat{\varphi}_N^m(\beta)$ *and* $\hat{\Omega}^m \xrightarrow{p} \Omega$, *which is positive definite. Let Assumptions A3.1-A3.4 hold. Suppose further that* $\mathcal{B}$ *is compact, that* $\varphi(z, v, \beta)$ *is continuous in* β *for every* z *and almost every* v, *that* $\mathbb{E}_{P_0} \max_z \sup_{\beta\in\mathcal{B}} ||\varphi(z, V, \beta)|| < \infty$, *and that* $\mathbb{E}_{P_0}\varphi(Z, V, \beta) = 0$ *if and only if* $\beta = \beta_0$. *Then* $\hat{\beta}^m \xrightarrow{p} \beta_0$ *for each* m *and* $\hat{\beta}_{MI} \xrightarrow{p} \beta_0$ *for any* M.

Proof. By Proposition A3.10, $\sup_{\beta\in\mathcal{B}} ||\hat{\varphi}_N^m(\beta) - \varphi_0(\beta)|| \xrightarrow{p} 0$ and $\varphi_0(\beta) = \mathbb{E}_{P_0}\varphi(Z, V, \beta)$ is continuous in β and therefore bounded on the compact set $\mathcal{B}$. From here, the proof follows the same logic as @Newey94 Theorem 2.6, and $\hat{\beta}^m \xrightarrow{p} \beta_0$. The average of a finite sum of values each converging in probability to β_0 also converges in probability to β_0, so $\hat{\beta}_{MI} \xrightarrow{p} \beta_0$. □

Continuous Plug-in Functionals

This section will prove the consistency of estimators which are functional of the empirical distribution of the data. Define the empirical distribution of the m-th imputed dataset as

$$\mathbb{P}_N^m = \frac{1}{N}\sum_{i=1}^N \delta_{(Z_i^m, V_i)}$$

Let $\mathcal{F}$ be a class of functions $f : \mathcal{Z} \times \mathbb{V} \to \mathbb{R}$ with $\int f \, d\mathbb{P}_N^m = \frac{1}{N}\sum_i f(Z_i^m, V_i)$.

Define a *bracket* $[l, u]$ as a pair of measurable functions on $\mathcal{Z} \times \mathbb{V}$ with $l \leq u$ everywhere and $\mathbb{E}_{P_0} \max_{z\in\mathcal{Z}}(|l(z, V)| + |u(z, V)|) < \infty$. Its *size* is $\int u - l \, dP_0 = \mathbb{E}_{P_0}[u(Z, V) - l(Z, V)]$. A function f *lies in* $[l, u]$ if $l(z, v) \leq f(z, v) \leq u(z, v)$ for every $z \in \mathcal{Z}$ and every $v \in \mathbb{V}$. A class $\mathcal{F}$ of measurable functions on $\mathcal{Z} \times \mathbb{V}$ satisfies the *bracketing condition* if, for every $\epsilon > 0$, there are finitely many brackets of size at most ϵ such that every $f \in \mathcal{F}$ lies in at least one of them.

These are very similar to the brackets used in the bracketing Glivenko-Cantelli theorem (van der Vaart, 1998, Theorem 19.4), with two exceptions. First, $l < u$ must hold everywhere,

even at values of z that are impossible at v under P_0. Second, l and u must be integrable in the sense that $\mathbb{E}_{P_0} \max_{z \in \mathcal{Z}}(|l(z,V)| + |u(z,V)|) < \infty$, rather than merely P_0-integrable as in van der Vaart (1998). Again, the difference is the maximum over z, which requires consideration of values of z impossible at v. These additional conditions hold for common constructions of brackets used in applications of the Glivenko-Cantelli theorem, such as for the class of indicator functions for continuous components Y of V: $\mathbf{1}(z = z_0)\mathbf{1}(y \leq t)$ for $z_0 \in \mathbb{Z}$ and $t \in \mathbb{R}$.

Proposition A3.13 (Glivenko-Cantelli theorem for imputed samples)**.** *Let Assumptions A3.1-A3.4 hold, and let $\mathcal{F}$ satisfy the bracketing condition. Then for each m,*

$$\sup_{f \in \mathcal{F}} \left| \int f \, d\mathbb{P}_N^m - \int f \, dP_0 \right| \xrightarrow{p} 0$$

Proof. Fix $\epsilon > 0$ and let $[l_1, u_1], \ldots, [l_J, u_J]$ be brackets of size at most ϵ such that every $f \in \mathcal{F}$ lies in (at least) one of them. If f lies in $[l_j, u_j]$, then $\int f \, d\mathbb{P}_N^m \leq \int u_j \, d\mathbb{P}_N^m$ because $\mathbb{P}_N^m$ is a probability distribution and $f \leq u_j$ at every point of $\mathcal{Z} \times \mathbb{V}$. Because $f \geq l_j$, $\int u_j - f \, dP_0 \leq \int u_j - l_j \, dP_0$. Therefore

$$\int f \, d\mathbb{P}_N^m - \int f \, dP_0 \leq \left[\int u_j \, d\mathbb{P}_N^m - \int u_j \, dP_0 \right] + \int u_j - f \, dP_0 \leq \left[\int u_j \, d\mathbb{P}_N^m - \int u_j \, dP_0 \right] + \epsilon$$

and symmetrically

$$\int f \, d\mathbb{P}_N^m - \int f \, dP_0 \geq \left[\int l_j \, d\mathbb{P}_N^m - \int l_j \, dP_0 \right] - \epsilon$$

Therefore,

$$\sup_{f \in \mathcal{F}} \left| \int f \, d\mathbb{P}_N^m - \int f \, dP_0 \right| \leq \max_j \max \left\{ \left| \int u_j \, d\mathbb{P}_N^m - \int u_j \, dP_0 \right|, \left| \int l_j \, d\mathbb{P}_N^m - \int l_j \, dP_0 \right| \right\} + \epsilon$$

and

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\int f\ d\mathbb{P}_N^m - \int f\ dP_0\right| > 2\epsilon\right) \leq$$

$$\mathbb{P}\left(\max_j \max\left\{\left|\int u_j\ d\mathbb{P}_N^m - \int u_j\ dP_0\right|, \left|\int l_j\ d\mathbb{P}_N^m - \int l_j\ dP_0\right|\right\} > \epsilon\right)$$

Each of the 2J terms inside the maximum $\xrightarrow{p} 0$ by Proposition A3.10 with $h = u_j$ and $h = l_j$. A maximum of finitely many terms that each $\xrightarrow{p} 0$ also does, so this completes the proof. □

Theorem A3.14 (Continuous plug-in functionals.)**.** *Let $\mathcal{F}$ satisfy the bracketing condition, and let $\mathcal{D}$ be a set of probability distributions on $\mathcal{Z}\times\mathbb{V}$ that contains P_0 and every distribution with finite support, such that for each $P\in\mathcal{D}$ every $f\in\mathcal{F}$ is P-integrable and* $\sup_{f\in\mathcal{F}}|\int f\ dP| < \infty$*. (The* sup *condition is satisfied for P_0 and for every finitely supported P, since for each f, $|f| \leq |l_j| + |u_j|$ for at least one j.) Let $\psi : \mathcal{D}\to\mathbb{R}^k$, and suppose that ψ is continuous at P_0 with respect to $d_{\mathcal{F}}(P, P_0) := \sup_{f\in\mathcal{F}}|\int f\ dP - \int f\ dP_0|$, i.e. for every $\epsilon > 0$, there is a $\delta > 0$ such that $||\psi(P) - \psi(P_0)|| < \epsilon$ whenever $P\in\mathcal{D}$ and $d_{\mathcal{F}}(P, P_0) < \delta$. Let $\hat{\beta}^m = \psi(\mathbb{P}_N^m) + o_p(1)$ and $\beta_0 = \psi(P_0)$. Under Assumptions A3.1-A3.4, $\hat{\beta}^m \xrightarrow{p} \beta_0$ for each m and $\hat{\beta}_{MI} \xrightarrow{p} \beta_0$ for any M.*

Proof. Fix $\epsilon > 0$ and choose δ as in the continuity condition. Since $\mathbb{P}_N^m$ has finite support, it belongs to $\mathcal{D}$ and

$$\mathbb{P}(||\psi(\mathbb{P}_N^m) - \psi(P_0)|| \geq \epsilon) \leq \mathbb{P}(d_{\mathcal{F}}(\mathbb{P}_N^m, P_0) \geq \delta) \xrightarrow{p} 0$$

by Proposition A3.13. The $o_p(1)$ difference does not affect the limit, and the average of finitely many terms that $\xrightarrow{p} \beta_0$ also converges to β_0. □

Ignorable Sampling Designs

The above results assumed a simple random validation sample. This section establishes that these results continue to hold under a wider class of sampling designs, described by the

following assumption.

Assumption A3.15 (Ignorable validation with positivity)**.** The triples (Z_i, V_i, R_i) are i.i.d. with $\mathbb{P}(R_i = 1 \mid Z_i, V_i) = \rho_N \pi(V_i)$. $\pi : \mathbb{V} \to (0, 1]$ does not depend on N, $\rho_N \in (0, 1]$, $N\rho_N \to \infty$, and $\sup_N \rho_N \mathbb{E}_{P_0} \pi(V) < 1$.

Lemma A3.16 (Change of measure)**.** *Let P^* and P be probability measures on $\mathbb{V}$ with P dominated by P^* and $w = dP/dP^*$. For every measurable $d : \mathbb{V} \to [0, 2]$ and every $K > 0$,*

$$\int d \, dP \leq 2\mathbb{P}(w > K) + K \int d \, dP^*$$

Therefore, if d_N is a sequence of random functions with values in $[0, 2]$ and $\int d_N \, dP^* \xrightarrow{p} 0$, then $\int d_N \, dP \xrightarrow{p} 0$.

Proof. $\int d \, dP = \int dw \, dP^* \leq 2 \int_{\{w>K\}} w \, dP^* + K \int d \, dP^* = 2\mathbb{P}(w > K) + K \int d \, dP^*$. Since $w < \infty$ almost surely, $\mathbb{P}(w > K) \to 0$ as $K \to \infty$.

Given $\epsilon > 0$, choose K with $2\mathbb{P}(w > K) < \epsilon$. Then $\mathbb{P}(\int d_N \, dP > 2\epsilon) \leq \mathbb{P}(K \int d_N \, dP^* > \epsilon) \to 0$. □

Assumption A3.17 (Imputation model consistency under ignorable sampling)**.** For each m, $\int d^m \, dP_V^* \xrightarrow{p} 0$.

P_V^*, the distribution of the validated V, is fixed $P_V^*(dv) = \frac{\pi(v) P_{0,V}(dv)}{\mathbb{E}_{P_0} \pi(V)}$, so Appendix 2 applies, justifying this assumption.

Corollary A3.18 (Ignorable validation designs)**.** *Suppose Assumptions A3.1, A3.3, A3.15, and A3.17 hold. Then Lemmas A3.5, A3.6, and A3.7 hold.*

Proof. (Lemma A3.5) Under Assumption A3.15, the triples (Z_i, V_i, R_i) are i.i.d. and R_i is independent of Z_i given V_i.

(Lemma A3.6) This holds by construction, exactly as in Lemma A3.6.

(Lemma A3.7) The non-validated $V_i \sim P(V \mid D_N^*, \mathcal{U}_S, R = 0)$. This conditional distribution's density with respect to $P_{0,V}$ is $\frac{1-\rho_N \pi(v)}{1-\rho_N \mathbb{E}_{P_0}\pi(V)} \leq \frac{1}{c}$, where $c = 1 - \sup_N \rho_N \mathbb{E}_{P_0}\pi(V) > 0$.

$$\mathbb{E}\left[\hat{\Delta}_N^m \mid D_N^*, \mathcal{U}_S\right] \leq \frac{1}{c} \int d^m \; dP_{0,V}$$

Since $\pi(v) > 0$, P_V^* dominates $P_{0,V}$ with density $w = \mathbb{E}_{P_0}\pi(V)/\pi(v)$. Lemma A3.16 and Assumption A3.17 therefore give $\int d^m \; dP_{0,V} \xrightarrow{p} 0$. The rest of Lemma A3.7 follows as before. □

As the consistency proofs throughout this appendix only make use of Assumptions A3.2 and A3.4 through these Lemmas, this suffices to generalize the consistency proofs to ignorable validation sampling designs.

References

Dennis J. Aigner. Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics*, 1(1):49–59, 1973. doi: 10.1016/0304-4076(73)90005-5.

Olanrewaju Akande, Fan Li, and Jerome Reiter. An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*, 71(2):162–170, 2017. doi: 10.1080/00031305.2016.1277158.

Jennifer Alix-Garcia and Daniel L. Millimet. Remotely incorrect? Accounting for nonclassical measurement error in satellite data on deforestation. *Journal of the Association of Environmental and Resource Economists*, 10(5):1335–1367, 2023. doi: 10.1086/723723.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000.

John Barnard and Donald B. Rubin. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4):948–955, 1999. doi: 10.1093/biomet/86.4.948.

Erich Battistin, Michele De Nadai, and Barbara Sianesi. Misreported schooling, multiple measures and returns to educational qualifications. *Journal of Econometrics*, 181(2): 136–150, 2014. doi: 10.1016/j.jeconom.2014.03.002.

Lane F. Burgette and Jerome P. Reiter. Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172(9):1070–1076, 2010. doi: 10.1093/aje/kwq260.

Rossella Calvi, Arthur Lewbel, and Denni Tommasi. LATE with missing or mismeasured treatment. *Journal of Business & Economic Statistics*, 40(4):1701–1717, 2022. doi: 10.1080/07350015.2021.1970573.

David Card. The effect of unions on the structure of wages: A longitudinal analysis. *Econometrica*, 64(4):957–979, 1996. doi: 10.2307/2171852.

Xiaohong Chen, Yingyao Hu, and Arthur Lewbel. A note on the closed-form identification of regression models with a mismeasured binary regressor. *Statistics & Probability Letters*, 78 (12):1473–1479, 2008a. doi: 10.1016/j.spl.2007.12.024.

Xiaohong Chen, Yingyao Hu, and Arthur Lewbel. Nonparametric identification of regression models containing a misclassified dichotomous regressor without instruments. *Economics Letters*, 100(3):381–384, 2008b. doi: 10.1016/j.econlet.2008.03.015.

Xiaohong Chen, Yingyao Hu, and Arthur Lewbel. Nonparametric identification and estimation of nonclassical errors-in-variables models without additional information. *Statistica Sinica*, 19(3):948–968, 2009.

Yi-Hau Chen and Hung Chen. A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(3):449–460, 2000. doi: 10.1111/1467-9868.00243.

Jef de Slegte, Filip Van Droogenbroeck, Bram Spruyt, Sam Verboven, and Vincent Ginis. The use of machine learning methods in political science: An in-depth literature review. *Political Studies Review*, 23(3):764–784, 2024. doi: 10.1177/14789299241265084.

Melissa Dell. Deep learning for economists. *Journal of Economic Literature*, 63(1):5–58, 2025. doi: 10.1257/jel.20241733.

Francis J. DiTraglia and Camilo García-Jimeno. Identifying the effect of a mis-classified, binary, endogenous regressor. *Journal of Econometrics*, 209(2):376–390, 2019. doi: 10.1016/j.jeconom.2019.01.007.

Dave Donaldson and Adam Storeygard. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–198, 2016. doi: 10.1257/jep.30.4.171.

David B. Dunson and Chuanhua Xing. A nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051, 2009. doi: 10.1198/jasa.2009.tm08439.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. Using Large Language Model annotations for the social sciences: A general framework for using predicted variables in downstream analyses. *Working Paper*, 2024.

Vicky Espinoza, Lorenzo Ade Booth, and Joshua H. Viers. Land use misclassification results in water use, economic value, and ghg emission discrepancies in california's high-intensity agriculture region. *Sustainability*, 15(8), 2023. doi: 10.3390/su15086829.

Christian Fong and Matthew Tyler. Machine learning predictions as regression covariates. *Political Analysis*, 29(4):467–484, 2021. doi: 10.1017/pan.2020.38.

Harley Frazis and Mark A. Loewenstein. Estimating linear regressions with mismeasured, possibly endogenous, binary explanatory variables. *Journal of Econometrics*, 117(1):151–178, 2003. doi: 10.1016/S0304-4076(03)00121-0.

Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992. doi: 10.1214/ss/1177011136.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, third edition, 2014.

Subhashis Ghosal and Aad van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Annals of Statistics*, 35(2):697–723, 2007. doi: 10.1214/009053606000001271.

Nick Hagerty. Transaction costs and the gains from trade in water markets. *Working Paper*, 2024.

J. A. Hausman, Jason Abrevaya, and J. M. Scott-Morton. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87(2):239–269, 1998. doi: 10.1016/S0304-4076(98)00015-3.

Yingyao Hu. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144(1):27–61, 2008. doi: 10.1016/j.jeconom.2007.12.001.

Hemant Ishwaran and Lancelot F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001. doi: 10.1198/016214501750332758.

Meha Jain. The benefits and pitfalls of using satellite data for causal inference. *Review of Environmental Economics and Policy*, 14(1):157–169, 2020. doi: 10.1093/reep/rez023.

A. Jasra, C. C. Holmes, and D. A. Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 2005. doi: 10.1214/088342305000000016.

Wenlong Ji, Lihua Lei, and Tijana Zrnic. Predictions as surrogates: Revisiting surrogate outcomes in the age of AI. *Working Paper*, 2025.

Hiroyuki Kasahara and Katsumi Shimotsu. Identification of regression models with a misclassified and endogenous binary regressor. *Econometric Theory*, 38(6):1117–1139, 2022. doi: 10.1017/S0266466621000451.

Michael P. Keane and Robert M. Sauer. Classification error in dynamic discrete choice models: Implications for female labor supply behavior. *Econometrica*, 77(3):975–991, 2009. doi: 10.3982/ECTA7642.

B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, 34(2):837–877, 2006. doi: 10.1214/009053606000000029.

Dan M. Kluger, Kerri Lu, Tijana Zrnic, Sherrie Wang, and Stephen Bates. Prediction-powered inference with imputed covariates and nonuniform sampling. *Working Paper*, 2025.

Brent Kreider. Regression coefficient identification decay in the presence of infrequent classification errors. *Review of Economics and Statistics*, 92(4):1017–1023, 2010. doi: 10.1162/REST_a_00044.

Arthur Lewbel. Estimation of average treatment effects with misclassification. *Econometrica*, 75(2):537–551, 2007. doi: 10.1111/j.1468-0262.2006.00756.x.

Kerri Lu, Dan M. Kluger, Stephen Bates, and Sherrie Wang. Regression coefficient estimation from remote sensing maps. *Remote Sensing of Environment*, 330, 2025. doi: 10.1016/j.rse.2025.114949.

Aprajit Mahajan. Identification and estimation of regression models with misclassification. *Econometrica*, 74(3):631–665, 2006. doi: 10.1111/j.1468-0262.2006.00677.x.

Xiao-Li Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4):538–558, 1994.

Bruce D. Meyer and Nikolas Mittag. Misclassification in binary choice models. *Journal of Econometrics*, 200(2):295–311, 2017. doi: 10.1016/j.jeconom.2017.06.012.

Jonas M. Mikhaeil, Andrew Gelman, and Philip Greengard. Hierarchical Bayesian models to mitigate systematic disparities in prediction with proxy outcomes. *Journal of the Royal Statistical Society Series A: Statistics in Society*, forthcoming. doi: 10.1093/jrsssa/qnae142.

Daniel L. Millimet. The elephant in the corner: A cautionary tale about measurement error in treatment effects models. *IZA Discussion Papers*, 5140, 2010.

Jared S. Murray. Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, 33(2):142–159, 2018. doi: 10.1214/18-STS644.

Jared S. Murray and Jerome P. Reiter. Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479, 2016. doi: 10.1080/01621459.2016.1174132.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In Robert F. Engle and Daniel L. McFadden, editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. 1994. doi: 10.1016/S1573-4412(05)80005-4.

Pierre Nguimkeu, Augustine Denteh, and Rusty Tchernis. On the estimation of treatment effects with endogenous misreporting. *Journal of Econometrics*, 208(2):487–506, 2019. doi: 10.1016/j.jeconom.2018.10.005.

Margaret Sullivan Pepe and Thomas R. Fleming. A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association*, 86(413): 108–113, 1991. doi: 10.1080/01621459.1991.10475009.

Jonathan Proctor, Tamma Carleton, and Sandy Sum. Parameter recovery using remotely sensed variables. *NBER Working Paper No. 30861*, 2023.

Ashesh Rambachan, Rahul Singh, and Davide Viviano. Program evaluation with remotely sensed outcomes. *Working Paper*, 2025.

Jerome P. Reiter and Trivellore E. Raghunathan. The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102(480):1462–1471, 2007. doi: 10.1198/016214507000000932.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987.

Donald B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996. doi: 10.2307/2291635.

Luke C. Sanford, Megan Ayers, Matthew Gordon, and Eliana Stone. Adversarial debiasing for unbiased parameter recovery. *Working Paper*, 2025.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2): 639–650, 1994.

Yajuan Si and Jerome P. Reiter. Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, 38(5):499–521, 2013. doi: 10.3102/1076998613480394.

Adrian L. Torchiana, Ed Rosenbaum, Paul T. Scott, and Eduardo Souza-Rodrigues. Improving estimates of transitions from satellite data: A hidden Markov model approach. *Review of Economics and Statistics*, 107(2):426–441, 2025. doi: 10.1162/rest_a_01301.

USDA NASS. Cropland data layers FAQs. 2024. URL https://www.nass.usda.gov/Research_and_Science/Cropland/sarsfaqs2.php.

A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Yuefang Wu and Subhashis Ghosal. The L_1-consistency of Dirichlet mixturies in multivariate Bayesian density estimation. *Journal of Multivariate Analysis*, 10(101):2411–2419, 2010. doi: 10.1016/j.jmva.2010.06.012.

David Wuepper, Wyclife Agumba Oluoch, and Hadi Hadi. Satellite data in agricultural and environmental economics: Theory and practice. *Agricultural Economics*, 56(3):493–511, 2025. doi: 10.1111/agec.70006.

Takahide Yanagi. Inference on local average treatment effects for misclassified treatment. *Econometric Reviews*, 38(8):938–960, 2019. doi: 10.1080/07474938.2018.1485833.