A Unifying Framework for Robust and Efficient Inference with Unstructured Data

Jacob Carlson and Melissa Dell*

March 20, 2025

Abstract

This paper presents a general framework for conducting efficient and robust inference on parameters derived from unstructured data, which include text, images, audio, and video. Economists have long incorporated data extracted from texts and images into their analyses, a practice that has accelerated with advancements in deep neural networks. However, neural networks do not generically produce unbiased predictions, potentially propagating bias to estimators that use their outputs. To address this challenge, we reframe inference with unstructured data as a missing structured data problem, where structured data are imputed from unstructured inputs using deep neural networks. This perspective allows us to apply classic results from semiparametric inference, yielding valid, efficient, and robust estimators based on unstructured data. We formalize this approach with MARS (**M**issing **A**t **R**andom **S**tructured Data), a unifying framework that integrates and extends existing methods for debiased inference using machine learning predictions, linking them to a variety of older, familiar problems such as causal inference. MARS makes it feasible to apply debiasing to a wide variety of settings by deriving estimators for common scenarios that have received little attention in the existing literature. We develop robust and efficient estimators for descriptive and causal estimands, address common empirical challenges such as inference with aggregated and (non-linearly) transformed structured data, and consider issues arising in large datasets. Finally, we reanalyze well-known studies that rely on unstructured data, demonstrating the practical impact of MARS.

1 Introduction

Economists frequently use unstructured data in empirical research. These include images (pixel data), text (token sequences from a high-dimensional vocabulary), audio (waveforms or spectrograms), and videos (image sequences). However, unstructured data are rarely used directly in econometric analyses, because they are high dimensional, computationally complex, and uninterpretable in their raw form. Instead, researchers extract meaningful, low-dimensional features from unstructured data and use these in statistical analyses. Following the literature, we refer to these low-dimensional representations as *structured data*.

*We have no funding sources or conflicts of interest to disclose.

Datasets of low-dimensional features extracted from unstructured data are fundamental to economic research across many fields. For example, commonly used datasets extract structured information on governance, institutions, political stability, policy uncertainty, conflict, and violence from news articles and other text data.[1] Researchers impute sentiment, topics, and a variety of other structured data from government transcripts, firm filings, earnings calls, patent data, and web texts.[2] Satellite images of nighttime lights are used to measure economic activity, development, and urbanization.[3] Imputations from remote sensing data frequently supplement ground measurements of temperature, precipitation, pollution, agricultural output, land use, illicit activities, deforestation, and other phenomena.[4]

Traditionally, extracting structured information from unstructured data was costly, requiring entirely manual annotation or complex, human-engineered rules. As a result, large-scale initiatives were often required. More recently, advances in computing and deep learning have significantly reduced the cost of extracting low-dimensional structured data from text, images, audio, and video. A vast machine learning literature demonstrates that deep neural networks are the state-of-the-art method for large-scale feature extraction from unstructured data (Goodfellow et al., 2016).

However, neural networks will not generically produce unbiased predictions in finite samples. Choices related to network architecture, the distribution of training data, and various implementation details can all introduce systematic biases. Moreover, the use of nonlinear transformations at each layer of the neural network and the frequent application of neural networks to binary or multiclass classification problems violate classical measurement error assumptions. Biases will propagate to estimators that rely on these predictions, affecting both point estimates and uncertainty quantification. In large datasets, while sampling variation may be small, a poorly performing first-step predictor can still introduce substantial uncertainty. Concerns about imputation bias are further heightened by the availability of off-the-shelf neural networks, which are relatively inexpensive to implement and can be inaccurate. Different neural networks may introduce different biases that then propagate to the estimators that rely on them, raising concerns that neural network-based imputations could be selectively used to produce desirable results.

As economists increasingly rely on neural networks to extract structured data for empirical analysis, a frequent question that arises is how accurate their predictions need to be. Improving neural network performance often involves significant costs, including training larger models, collecting more or higher-quality training data, and refining complex implementation details. To ensure unbiased estimates and determine whether costly investments to improve first step predictions are necessary, researchers need a framework that explicitly accounts for first-step imputation error when using neural network predictions.

To address these challenges, this paper introduces MARS (***M**issing **A**t **R**andom **S**tructured*

[1]For example, the Economic Policy Uncertainty Index, the Armed Conflict Dataset, the Conflict Recurrence Dataset, the Correlates of War, the Global Terrorism Database, the Polity5 Dataset, Internal Wars & Failures of Governance, the Mass Mobilization Data, the Armed Conflict Location & Event Data, and the Geopolitical Risk Index.

[2]For example, FOMC Transcripts, various sources for earnings calls, USPTO and Google patent data, and web texts a from platforms like Reddit or Twitter.

[3]See Gibson et al. (2020) for a review.

[4]These include datasets on temperature and precipitation (*e.g.,* MODIS, ERA5, CHIRPS, TRMM), agriculture and land use (Harvest, Landsat, Sentinel),air quality and atmospheric composition (*e.g.,* OMI/Aura and TROPOMI/Sentinel-5P, MERRA-2, CAMS), and deforestation and illicit activity (*e.g.,* PlanetScope,UNODC-ICMP)

Data), a framework for conducting *valid, efficient*, and *robust* inference on estimands that incorporate unstructured data through their low-dimensional features. MARS frames inference with unstructured data as a missing data problem, because raw unstructured datasets commonly lack the low-dimensional summaries that are relevant to economic analyses. The framework builds upon fundamental work on missing data, in particular the Rubin (1976) missing at random (MAR) mechanism (see also Little and Rubin (2019); Robins et al. (1994, 1995); Robins and Rotnitzky (1995); Bang and Robins (2005)). MARS leverages foundational results from semiparametric inference with missing data, which provides a well-established, broadly applicable, and assumption-light approach to debiasing estimators. The key idea is to collect a validation sample containing ground truth feature values, use this sample to estimate the bias in the imputed data, and adjust estimates accordingly.

MARS requires access to ground truth data. This requires explicitly defining the features to be extracted from unstructured data and establishing a clear, implementable procedure for measuring them accurately. Ground truth is obtained through a costly process such as annotation of texts or images by highly skilled and motivated human experts or the collection of ground station data in the context of remote sensing. While deep neural networks are often treated as black boxes, MARS requires that their outputs be interpretable through the annotation process. Under MARS, the validation sample must meet the 'missing at random' assumption. This means that after adjusting for observable factors, annotated and unannotated structured data should be comparable in their ground truth values. This parallels the 'selection on observables' assumption in causal inference—a related missing data problem where control outcomes are missing for treated units and vice versa. In MARS, observations are selected into annotation (rather than treatment).

Debiasing with a ground truth annotation sample is central to recent influential frameworks for valid statistical inference using predictions from black box AI models (e.g., Angelopoulos et al. (2023); Egami et al. (2023); Ludwig et al. (2024)). The contribution of MARS to this literature is threefold: 1) developing a theoretical framework that unifies this work and links it to a variety of much older, familiar problems, 2) identifying estimators that are both unbiased and efficient, 3) making it feasible to apply debiasing to a wide variety of settings by deriving estimators for common scenarios that have received little attention in the existing literature.

First, MARS provides a theoretical framework that unifies recent work on inference with black-box AI models—developed independently across disciplines with limited interactions—and connects this work to a longstanding econometrics literature on measurement error (*e.g.,* Schennach (2016); Chen et al. (2005)), widely used inference methods that incorporate machine learning-based first steps (*e.g.,* Chernozhukov et al. (2018, 2022a,b)), and classical literatures on missing data and causal inference (*e.g.,* Rubin (1978); Imbens and Rubin (2015); Robins et al. (1994) and many others). MARS improves our understanding of inference with unstructured data by connecting it to familiar problems, such as causal inference. Viewing inference with unstructured data and causal inference as special cases of a more general missing data problem highlights various relevant insights from semiparametric inference.

We depart from existing work on black box AI by emphasizing a semiparametric approach, which provides new insights about efficiency. For example, for an estimator to achieve asymptotic efficiency, the imputation of missing structured data should depend not only on unstructured data (e.g., texts or images) but also on context-specific structured variables that help estimate

the target parameter (e.g., other covariates in a regression model). Some existing work on inference with black box AI argues that seimparametric methods are too complicated to be widely practical (Angelopoulos et al., 2024). While potentially a valid concern for some disciplines, in economics the opposite is likely to be true, as semiparametric methods have a long history of wide use and allow linking inference with unstructured data to more familiar problems like causal inference and debiased machine learning.

Our final set of contributions addresses unbiased and efficient inference with unstructured data in a variety of common scenarios in applied economics that have received little attention in the existing literature. We derive robust and efficient estimators for descriptive moments, linear regression model coefficients, treatment effects identified through linear IV models, modern difference-in-differences estimands, and regression discontinuity estimands under local randomization. Because it employs a semiparametric approach, MARS could also be integrated with recent advances in automatic debiasing (Chernozhukov et al., 2022b; van der Laan et al., 2025) or automatic (functional) differentiation (Luedtke, 2024) to further extend its applicability.

Existing de-biasing approaches assume that ground truth data are available at the same level of aggregation as the parameter of interest. However, economic analyses often require aggregating structured data, and non-linear aggregations are common. For example, researchers may measure conflict, (logged) policy uncertainty, or other characteristics at the year $\times$ geographic region level by aggregating imputations from a large volume of daily newspaper articles. Collecting (complete) ground truth data at this aggregate level would often require labeling thousands or even millions of unstructured data instances just to obtain a single ground truth observation. To address this challenge, we develop a framework that leverages MARS for debiased inference when ground truth data are available only at a disaggregated level, while the parameter of interest is derived from structured data that are aggregated.

When working with large datasets, the structured data of interest often represent a "rare event." For example, a text that pertains to a specific topic can be relatively rare compared to a corpus's size. The typical approach to this challenge, which arises frequently in empirical economics, is to only annotate data that meet some filter, such as containing certain keywords. Unless the population of interest is only instances that meet the filter, this violates the assumption that the annotation function not place zero probability on annotating certain types of observations. This is equivalent to a strong overlap assumption in causal inference. When researchers are annotating class-imbalanced structured data, the rare event estimation literature suggests optimizing the annotation function to reduce variance, potentially by incorporating importance sampling techniques (Sigman, 2010). We refer readers to a machine learning literature that develops this approach to annotation (Zrnic and Candes, 2024).

Together, these contributions substantially broaden the practical applicability of debiasing in applied economics settings. To illustrate how MARS can be deployed, we apply it to recent studies that use low-dimensional summaries of unstructured data, demonstrating that non-MARS estimates can produce overly precise confidence intervals and systematically differ from bias-corrected estimates. We also illustrate how it can be used with aggregated and (non-linearly) transformed structured data, which arise frequently in practice. While MARS is particularly relevant for deep neural networks, our analysis shows that it is equally applicable to other methods—such as sparse (keyword-based) approaches—that extract low-dimensional features from unstructured data. We also explore how various design choices within MARS, including

annotation budgets and the selection of deep neural networks, influence estimates.

A skeptic might raise several questions: Must annotated data be missing-at-random (conditional on observables)? Is defining a ground truth meaningful when the underlying concept is complex? And how much of a concern is imputation bias to begin with?

Ensuring that annotated data are missing at random enables double robustness—intuitively, weaker assumptions can be made about the imputation function when the annotation process helps achieve unbiased inference. While the missing-at-random assumption places restrictions on the annotation function, these restrictions are often feasible because researchers either annotate their own data or have access to a representative sample of ground truth. By imposing these conditions on annotation, MARS minimizes restrictions on the imputation function. This is particularly valuable when working with deep neural networks whose biases shift in a potentially complex way with input distributions and implementation details.

In some cases, imputations from unstructured data are necessary precisely because ground truth for the target population is unavailable, violating the missing-at-random assumption. Since MARS treats the neural network as a black box, the absence of ground truth for certain observations prevents measurement of imputation bias, making unbiased estimation impossible. Departures from missing at random could be addressed with additional restrictions on the imputation function—such as assuming a stable relationship between imputed and ground truth data across annotated and unannotated samples (see Rambachan et al. (2024a) for a framework using remote sensing data where ground truth is unavailable for the sample of interest).

While stronger assumptions about the imputation function are necessary in some contexts, researchers should be cautious when extrapolating estimates of imputation bias to samples that systematically differ from observations with available ground truth. A fundamental principle in machine learning is that predictive accuracy deteriorates when there is domain (covariate) shift away from the training data distribution (Ben-David et al. (2010)). Although there has been debate around whether very large models can reason beyond their training data distribution, many real-world examples—even with the largest models—demonstrate that performance degrades when neural networks are applied out-of-distribution. Neural networks are typically trained on a subset of annotated data or - in the case of off-the-shelf models - on easily accessible data that often differs from observations where ground truth is costly to obtain. MARS directly addresses this fundamental limitation of neural networks by imposing minimal assumptions on the imputation function and ensuring unbiased and robust estimates through the requirement that annotation meets the missing at random requirements.

A broader concern is that worry about imputation bias is unnecessary, since economists often aim to capture concepts that lack a uniquely definable ground truth, such as institutional quality or economic and geopolitical uncertainty. According to this view, imputations serve as proxies rather than precise measurements, making the notion of a well-defined ground truth irrelevant. However, even when the underlying concept is complex, explicitly defining the intended measurement and quantifying how well actual measurements align with it provide important scientific benefits. This approach compels researchers to clearly articulate their measurement objectives and evaluate how well their data reflect those objectives. The MARS framework ensures that the researcher can at least produce unbiased estimates based on their chosen definition.

A skeptic might also question whether first-step imputation error is a serious concern, given that economists have long treated structured variables extracted from unstructured data as

ground truth. Traditionally, off-the-shelf datasets lacked validation samples, leaving researchers with little choice but to accept them as (potentially biased) proxies. However, deep learning and evolving scientific norms are reshaping this landscape. Researchers now frequently generate their own large-scale structured datasets using deep neural networks, allowing them to design data collection for unbiased and efficient inference. Correcting for bias by annotating a small auxiliary sample of raw texts or images often requires only a modest investment, yet yields substantial benefits by ensuring robust, unbiased, and efficient inference.

This paper is organized as follows. Section 2 situates MARS within the relevant literature. Section 3 introduces the MARS framework, defining a model of missing structured data, outlining assumptions, establishing principles of robustness and efficiency, and presenting a general algorithm for applying MARS. Section 4 develops robust and efficient estimators for functionals commonly used in empirical economics. Section 5 explores connections between MARS and existing frameworks for valid inference using predictions from machine learning models. Section 6 extends MARS to address statistical challenges associated with large unstructured datasets. Section 7 demonstrates empirical applications of MARS in economic research, and Section 8 concludes.

2 Literature

Missing data. There is a well-established literature on missing data, pioneered in statistics and biostatistics (e.g., Rubin (1976); Little and Rubin (2019); Robins et al. (1994, 1995); Robins and Rotnitzky (1995); Bang and Robins (2005)). We draw on these canonical models, particularly the "missing at random" (MAR) mechanism, as an organizing framework (Rubin, 1976). Causal inference can itself be framed as a missing data problem, where treatment outcomes are missing for control observations and vice versa (Little and Rubin, 2019; Ding and Li, 2018; Hirano et al., 2003)). Consequently, missing data frameworks are widely used in causal inference (Imbens and Rubin, 2015) and econometrics (Athey et al., 2019), and the MARS notation and framework share key conceptual similarities with causal inference. Insights from causal inference help clarify important aspects of inference with unstructured data, as we highlight at several points.

Semiparametric inference. There is a vast body of research on semiparametric and nonparametric inference (e.g., Pfanzagl (1982); Bickel et al. (1998); Newey (1994); van der Vaart (1998); Kennedy (2016, 2018)). Semiparametric inference frameworks are flexible, making few substantive assumptions about the data-generating process (DGP), and letting the data speak for itself as much as possible when estimating parameters of interest. This makes them particularly suitable for missing data problems, as further explored in Tsiatis (2006). These frameworks are supported by well-established theories of minimax-style efficiency, which provide a benchmark for comparing the performance of different estimators (Newey, 1994; van der Vaart, 1998).

The augmented inverse propensity weighted (AIPW) estimator is a widely used semiparametric method for estimating treatment effects (Robins et al., 1994; Robins and Rotnitzky, 1995; Scharfstein et al., 1999). It belongs to a broader class of doubly robust estimators, which relax requirements on estimators of nuisance parameters. This form of robustness is a key property of the estimators considered in this paper, and the familiar structure of the AIPW estimator

appears in the examples discussed.

Advances in machine learning for function approximation have spurred influential research on semiparametric methods for inference on functionals with high-dimensional nuisance parameters estimated using machine learning models. This literature emphasizes inference on low-dimensional parameters based on nonparametric "first-step" estimation problems, particularly in the context of causal inference (*e.g.,* Chernozhukov et al. (2018, 2022a,b); Ichimura and Newey (2022)). Building on this tradition, we develop efficient and robust estimators for missing data functionals under assumptions tailored to unstructured data: deep neural networks as the most commonly used function class, high-dimensional covariates, and known annotation (propensity) scores.

Measurement error. Bias in features imputed from unstructured data can be thought of as a form of systematic measurement error, a topic extensively studied in econometrics. For recent comprehensive reviews, see (Chen et al., 2011; Schennach, 2016, 2022). A key insight from this literature is that validation samples with ground truth values provide a general, model-agnostic method for correcting non-classical measurement error in nonlinear models (Chen et al., 2005).

Debiased inference with predictions from black-box machine learning models. Recent research has focused on developing general-purpose frameworks for asymptotically valid inference with flexible function approximators, particularly black-box AI such as large language models (LLMs). Notable contributions in this area include "prediction-powered inference" (Angelopoulos et al., 2023, 2024; Zrnic and Candès, 2024; Zrnic and Candes, 2024; Ji et al., 2025; Kluger et al., 2025), "design-based supervised learning" (Egami et al., 2023, 2024), "flexible regression adjustment" (List et al., 2024), and the applied econometric framework of Ludwig et al. (2024). These works depart from semiparametric theory, in stark contrast to Chernozhukov et al. (2018), which centers it. By combining semiparametric methods with black box imputation from unstructured data, we show that these methods can be viewed as special cases unified under the banner of semiparametric inference for missing data. Semiparametric theory also brings insights into efficiency and draws parallels with problems economists are likely to be familiar with (*e.g.*, causal inference).

Semi-supervised inference. A literature in statistics and computer science on semi-supervised learning and semi-supervised inference (*e.g.*, Chapelle et al. (2010)) explores settings where the researcher has access to a dataset in which some data points have supervisory signals (*e.g.*, are labeled) and other data points do not (*e.g.*, are unlabeled). This literature examines how to combine labeled and unlabeled data for optimal learning or inference (Zhang and Bradic, 2022; Zhang et al., 2023; Kallus and Mao, 2024). This paper focuses on intricacies associated with unstructured data and the models suited for learning from them, and develops a variety of missing data functionals of interest to economists and other social scientists, which have not been a focus in this literature.

3 Missing At Random Structured Data

This section introduces the MARS framework ("**M**issing **A**t **R**andom **S**tructured data"). The name highlights that the framework builds upon the foundational missing-data mechanisms of

Rubin (1976). We start by establishing a terminology and model for data missingness in the context of unstructured data, then outline the key assumptions of MARS, and finish with a recipe for its practical application.

3.1 Modeling Missing Structured Data

To enable robust and efficient inference with *unstructured data*, we recast the problem as inference on *missing structured data*. Structured data, denoted $M \in \mathcal{M}$, are low-dimensional data that can be used directly in estimating equations. They contrast with unstructured data, denoted $U \in \mathcal{U}$, which are high-dimensional and unsuitable for direct use in estimation (e.g., raw RGB values of pixels, audio waveforms, or sparse term vectors).

Under MARS, structured data are observed through a process termed annotation. An annotator (a human expert or a costly measurement technology) records the M that corresponds to a given U. For example, a researcher may examine text to determine its topic or sentiment or label objects in an image. Alternatively, structured data might come from a trusted but costly measurement process, such as ground stations recording temperature or pollution levels, or a high-quality AI system, whose output is considered by the research community to define the ground truth.

Since ground truth annotation is too expensive to scale, researchers instead learn a function to impute missing structured data, allowing them to leverage the full unstructured dataset—often vastly larger than the manually labeled subset. Deep neural networks typically serve as the imputation function $\hat{\mu}$, because they are the state-of-the-art method for mapping unstructured data to low-dimensional outputs. However, while state-of-the-art, neural network imputations are rarely perfect, and their accuracy can vary significantly depending on the model and task complexity.

MARS, and the foundational missing data framework that it builds upon, are closely linked to the Rubin Causal Model (Neyman, 1923; Rubin, 1974, 1978; Imbens and Rubin, 2015), as causal inference is fundamentally a missing data problem. Thus, we incorporate the notation of potential outcomes into MARS. This provides a unified notation for discussing unstructured data (viewed as missing structured data), alongside other common forms of data missingness, such as those arising from the fundamental problem of causal inference.

Suppose that structured data $M \in \mathcal{M}$ are subject to some data missingness given by an annotation indicator variable $A \in \{0, 1\}$. A consistency of potential outcomes assumption (as in the Rubin Causal Model) then allows us to write:

$$M = A\, M^{a=1} + (1 - A)\, M^{a=0},$$

where $M^{a=1}$ is the "ground truth" potential outcome, observed when the researcher has labeled that unstructured data instance, and $M^{a=0}$ is the "missing" potential outcome, obtained in the state where the instance has not been annotated. We keep the argument a explicit for now, to distinguish it from potential outcomes that depend on other random variables, such as treatment status.

As in causal inference, one can view the probability distribution of the random variable M^a as the distribution of M in a counterfactual world where $A = a$ with probability 1, *e.g.,* the distribution of structured data when all data have been annotated. The researcher is interested

in functionals of this distribution.

Without loss of generality, we can set $M^{a=0} = 0$ with probability 1, since, given an assumption on the consistency of potential outcomes, we only need to enforce that $M^{a=0}$ carries no information about $M^{a=1}$. Under this standardization, we can write $M = A\,M^{a=1}$. For brevity of notation, we can relabel the ground truth potential outcome as $M^* := M^{a=1}$. As such, we can write:

$$M = A\,M^*.$$

This notation is familiar from Rubin's canonical missing data models, which assume that a researcher observes the (iid) dataset $\{A_i\,M_i^*\}_{i=1}^n$ and is interested in inference on M^*. This notation increases the legibility of writing consistency assumptions under multiple types of missingness (*e.g.,* structured data and causal inference) that can arise in a given setting, as will be explored later.

Often, the researcher will begin with an unstructured dataset (e.g., a database of texts or images) where all structured data are missing. Just as we cannot perform valid causal inference without assigning some units to treatment, we cannot perform valid inference on $M^{a=1}$ without annotating some data, ensuring that $P(A = 1) > 0$. Collecting ground truth values M^* focuses the researcher on defining a precise and implementable definition of the structured data. Conducting annotation in practice can raise non-trivial challenges, which we return to in Section 6.1.

Some concepts that social scientists seek to measure are complex and open to multiple interpretations. However, for MARS to be applicable, the researcher must define a clear and implementable standard for the concept they aim to capture. MARS does not apply in settings where the goal is to organize data meaningfully across multiple possible interpretations, without assessing the accuracy of imputation against a ground truth. While such an approach is appropriate in some contexts, explicitly defining and measuring structured data using a precise and reproducible standard offers important scientific benefits. More broadly, MARS is not suited for unsupervised settings, where the aim is to uncover patterns in unlabeled data. For econometric approaches to unstructured data without a defined ground truth, see, for example, Battaglia et al. (2024); Rambachan et al. (2024b).

Beyond being designed for supervised settings, MARS is a superpopulation-based framework that assumes data are independently and identically distributed (iid). Extensions to other inferential frameworks are left to future research.

3.2 Assumptions

We now state the four assumptions that underlie the MARS framework. An advantage of a semiparametric approach is that these assumptions place relatively few restrictions on the data generating process.

Assumption 1 (Consistency of potential outcomes)**.** *For ground truth potential outcome $M^* \in \mathcal{M}$, structured data $M \in \mathcal{M} \times \{0\}$, and indicator $A \in \{0, 1\}$, we have*

$$M = AM^*.$$

In causal inference, consistency of potential outcomes requires that the treatment is well-

defined and that each observation's outcome depends solely on its own treatment status (non-interference). In settings with one-sided data missingness, the concept relies on analogous principles. Annotation status must be well-defined—each observation is either labeled or not—and the ground truth label for any given instance should depend only on its own annotation status, not on the annotation status of other instances. This is ensured by using a precise definition for annotation that remains unchanged throughout the annotation process.

The second assumption is the cornerstone of the MARS framework. It states that Rubin's "missing at random" (MAR) assumption holds for structured data (the ground truth potential outcomes), given observables the researcher has access to (Rubin, 1976; Little and Rubin, 2019).

Assumption 2 (Missing at random structured data)**.** *For ground truth potential outcome $M^* \in \mathcal{M}$, annotation indicator $A \in \{0, 1\}$, observed covariates $X \in \mathcal{X}$, and unstructured data $U \in \mathcal{U}$:*

$$[(U, M^*) \perp\!\!\!\perp A] \mid X.$$

Drawing an analogy to the selection on observables assumption in causal inference (Hirano et al., 2003), Assumption 2 could also be labeled "annotation on observables," as it states that after adjusting for observables X, annotated and unannotated structured data (and the associated unstructured data) are comparable in their ground truth values. There are no unaccounted for confounders determining whether an instance of unstructured data is annotated.

When researchers create their own structured datasets - common in the era of deep learning - they will often choose how to annotate data, and can design the annotation process to satisfy Assumption 2. In some scenarios, A may be completely randomized, $(U, M^*) \perp\!\!\!\perp A$, in which case Assumption 2 is guaranteed to hold.

The third assumption is that the annotation mechanism is known and can be bounded such that no instances of unstructured data are guaranteed to be either annotated or unannotated.

Assumption 3 (Known and bounded annotation score function)**.** *We define the "annotation score function" to be $\pi(x) := P(A = 1 \mid X = x)$. We assume that $\pi(x)$ is fixed, known, and bounded away from zero and one, i.e., $\pi(x) \in [\eta, 1 - \eta]$ for $0 < \eta \leq 1 - \eta < 1$ and for all $x \in \mathcal{X}$.*

We choose the naming convention "annotation score function" to mimic the terminology of the propensity score function.

In cases where the researcher does not annotate their own structured data, the annotation function may need to be estimated. The assumption that the annotation function is known can be relaxed, provided the researcher is willing to impose rate requirements on the imputation function and the (estimated) annotation function. We elaborate on these requirements when discussing the final MARS assumption, which concerns the imputation function. In this case, instead of Assumption 3, we have:

Assumption 4 (Bounded annotation score function)**.** *We define the "annotation score function" to be $\pi(x) := P(A = 1 \mid X = x)$. We assume that $\pi(x)$ is bounded away from zero and one, i.e., $\pi(x) \in [\eta, 1 - \eta]$ for $0 < \eta \leq 1 - \eta < 1$ and for all $x \in \mathcal{X}$.*

When Assumption 4 holds but the annotation function must be estimated, the statistical model under which inference is performed becomes fully nonparametric: $P \in \mathcal{P}$, as opposed to $P \in \mathcal{P}_\pi$. However, for any MARS mean functional, we will show subsequently that the efficient influence function remains the same under $\mathcal{P}_\pi$ or $\mathcal{P}$; as such, all EIFs derived for MARS mean

functionals, including those in the application section, remain the same when Assumption 4 replaces Assumption 3. Hence, the constructions of efficient estimators also remains the same.

Whether under assumption 3 or 4, the annotation score function must be bounded, which embeds the assumption of "strong overlap" commonly used in observational causal inference. In such settings, the plausibility of strong overlap diminishes as the dimensionality of the variable ensuring unconfoundedness increases (D'Amour et al., 2021). However, in most economic applications with unstructured data, X is relatively low-dimensional, whereas U is often high-dimensional. Notably, the machine learning literature on annotation sometimes specifies X as a low-dimensional representation of U, a perspective we will revisit when discussing annotation in practice with MARS in Section 6.1.

Existing annotation practice in social science often violates strong overlap. When working with text data, social scientists often rely on keyword-based filtering, assigning a nonzero probability of annotation to texts containing specific keywords while assigning a zero probability to all other unstructured data instances. Intuitively, the measurement error of a language model is likely to be systematically correlated with the terms appearing in the text. One way to address this issue is by redefining the population of interest to include only instances containing the relevant keywords. However, this alters the interpretation of the resulting estimands and estimates. Fortunately, more robust alternatives to keyword-based methods exist, which assign positive probability to sampling all unstructured data. These approaches are discussed in detail in Section 6.1.

A similar issue arises when annotation data are collected for one time period and then extrapolated forward or backward to other periods. If the data-generating process changes over time, this can introduce systematic measurement error correlated with time. For example, distinguishing crop types may become more challenging due to shifts in farming practices, climate, or the crops themselves (Wang et al., 2020).

MARS is not applicable in settings where imputation is used because collecting ground truth data is impossible for certain observations that are systematically different from those with available ground truth. For example, researchers may rely on imputations from satellite data to study regions where conflict or extreme isolation prevents ground data collection, while annotation data are entirely drawn from politically stable and accessible regions.

The final core assumption of MARS concerns the function for imputing missing structured data:

$$\mu(\tilde{x}) := E[M \mid A = 1, \tilde{X} = \tilde{x}],$$

where $\tilde{X} := (X, U, Q)$. Q is a set of context specific random variables that influence the estimation of our quantity of interest, such as controls in an ordinary least squares regression. As we will see in the following section, these may be necessary for efficient imputation.

In practice, μ is unknown and must be approximated. We refer to this estimator $\hat{\mu}$ as the "imputation function." Accurate predictions from the imputation function capture meaningful information about the ground truth distribution of structured data and improve precision when the predictions are used to estimate a functional of interest. Here, as in other semiparametric causal inference settings, we make an assumption about the quality of an estimator $\hat{\mu}$ for μ in the large sample limit.

Assumption 5 (MSE consistency of the imputation function)**.** *For function $\mu(\tilde{x})$, estimated*

conditional expectation function ("imputation function") $\hat{\mu}(\tilde{x})$, and features $\tilde{X} \in \tilde{\mathcal{X}}$, we have that

$$E\left[\left(\hat{\mu}(\tilde{X}) - \mu(\tilde{X})\right)^2\right] = o(1).$$

Intuitively, this condition says that we need the expected square error of our estimator to go to zero as the amount of data we train the estimator with goes to infinity; in other words, the estimator is well-specified. Assumption 5, sometimes referred to as "universal consistency" (Wager, 2024), is very mild in the context of deep neural networks. Recent theoretical work has shown that certain classes of deep neural networks learned with gradient descent—as will be deployed with MARS—are universally consistent (Drews and Kohler, 2024).

When Assumption 4 replaces Assumption 3, we intuitively lose some robustness in our estimators: we must now estimate the annotation score function, and Assumption 5 (universal consistency of the imputation function) must be replaced with a stronger set of assumptions on (MSE or $L^2(P)$) convergence rates for both the imputation function and the estimated annotation score function in order to guarantee efficient, $\sqrt{n}$-consistent, and asymptotically normal inference. In particular, the researcher now needs to assume that the convergence of the imputation and estimated annotation score functions occurs at a sufficiently fast rate. For discussion of these rates, one can consult a host of results from the semiparametric causal inference literature (Chernozhukov et al., 2018; Kennedy, 2023; Wager, 2024), e.g., it is sufficient that both the imputation and estimated annotation score functions converge at order $n^{-1/4}$ rates.

Classic semiparametric theory yields "curse of dimensionality" results for learning high-dimensional (nonparametric) conditional expectation functions, such as $\mu : \tilde{X} \to \mathbb{R}$, implying that minimax rates for convergence of the imputation function can be quite poor for large $k + \ell$. However, in practice, nonparametric regression using deep neural networks on high-dimensional unstructured data appears to exhibit fast rates of convergence under common data-generating processes (DGPs) (Klaassen et al., 2024), likely due to the massive-scale datasets used for pre-training. For example, state-of-the-art LLMs are pre-trained on trillions of tokens.

3.3 Efficient and Robust Inference with Unstructured Data

A key advantage of the MARS framework is its ability to enable valid, efficient, and robust inference in unstructured data settings commonly encountered by economists.

A semiparametrically efficient estimator of a functional achieves the lowest possible asymptotic variance among all regular, asymptotically linear (RAL) estimators of that functional (for a formal statement, see, e.g., Theorem 25.20 of van der Vaart (1998)). In other words, an efficient estimator attains the semiparametric variance lower bound, which serves as a benchmark for optimality within this framework. This lower bound corresponds to the variance of the efficient influence function (EIF) associated with the functional. We denote the EIF by $\varphi_\theta(W_i; P)$ for a data instance $W_i \in \mathcal{W}$, a joint distribution $P \in \mathcal{P}$ where $W_i \overset{\text{iid}}{\sim} P$, and a functional $\theta \in \Theta \subseteq \mathbb{R}$.[5] Simplifying the notation to $\varphi(W_i)$, an "optimal" RAL estimator $\hat{\theta}$ satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \text{Var}(\varphi(W_i))).$$

[5]Efficiency is also defined for vector-valued parameters $\theta \in \Theta \subseteq \mathbb{R}^p$, though, for simplicity, we consider scalar-valued functionals in this paper.

Intuitively, an influence function captures how a small change in the data distribution impacts the value of a functional (*e.g.,* the mean, or other functionals of the data), or the probability limit of an estimator of a functional. Estimators are linked to influence functions through their probability limits (Newey, 1994; Ichimura and Newey, 2022), and functionals themselves are also associated with influence functions, sometimes referred to as "influence curves" (Kennedy, 2023; Hines et al., 2022).

In fully nonparametric statistical models—which is to say there are no restrictions placed on $\mathcal{P}$—any RAL estimator is necessarily efficient, as there is only one valid influence function for all RAL estimators, which is the efficient influence function (Chen and Santos, 2018; Kennedy, 2023; Schuler and van der Laan, 2024). However, in semiparametric models, multiple influence functions may exist for different RAL estimators, but only one achieves the semiparametric efficiency bound and corresponds to the EIF. Because the annotation score function π is known in the MARS framework, the statistical model is semiparametric, which we label $\mathcal{P}_\pi$. As a result, any RAL estimator is not automatically efficient. Multiple influence functions may exist, and the efficient influence function is the one with the lowest variance.

Robust estimation, in the context of semiparametric inference, focuses on constructing estimators that relax the rate requirements for first-step (or, nuisance) parameter estimates in multi-step procedures. The first-step/nuisance parameter estimators are not directly used to estimate the primary parameter of interest (e.g., a causal effect or functional) but are essential for constructing the final estimator (e.g., by imputing structured data).

Intuitively, robust estimators are designed to tolerate trade-offs in estimation error among their first-step components while maintaining asymptotic normality and $\sqrt{n}$-consistency. In such estimators, worse performance in one component can be offset by better performance in another. An estimator is doubly robust when there are two nuisance functions to be estimated, and the estimator can balance errors between them without losing $\sqrt{n}$-consistency.

In the MARS framework, Assumptions 3 (known and bounded annotation score function) and 5 (MSE consistency of the imputation function) characterize robustness. The first-step estimator $\hat{\mu}$ (e.g., a deep neural network) is subject to weak conditions because we have access to the most accurate possible first-step estimator for the annotation function π, which is π itself. Intuitively, knowing the annotation score function significantly enhances the robustness of semiparametric estimation in MARS, exactly like how knowing the propensity score function would strengthen the robustness of semiparametric causal inference.

As previously mentioned, the intuition behind constructing robust and efficient estimators within the MARS framework is broadly similar to approaches in the semiparametric inference literature for missing data, including causal inference: impute the missing data using a sufficiently accurate first-step imputation function, and debias using ground truth values collected through annotation. However, several factors differentiate efficient and robust semiparametric estimation for unstructured data from other contexts.

First, *imputation functions $\hat{\mu}$ are typically deep neural networks.* Deep neural networks are state-of-the-art for predicting features from unstructured data. Neural networks such as transformers (Vaswani et al., 2017) are often needed to ensure that $\mu \approx \hat{\mu}$ either asymptotically or in finite samples. Although the MSE convergence rates for these models are unknown, Assumptions 3 (known and bounded annotation score function) and 5 (MSE consistency of the imputation function) imply that we do not need to focus on convergence rates but rather on

universal consistency, which theory indicates is achievable for deep neural networks (Drews and Kohler, 2024). Furthermore, a massive empirical literature demonstrates the success of deep neural networks in prediction tasks involving unstructured data, suggesting that convergence occurs quickly in practice (Klaassen et al., 2024). This is likely due to the widespread use of the "pre-train and fine-tune" transfer learning approach, which effectively increases the size of the dataset used to fit the imputation function (Cai and Pu, 2024).

Second, *features $\tilde{X}$ are high-dimensional.* This is because $U \in \tilde{X}$, and unstructured data U belong to the set $\mathcal{U} \subseteq \mathbb{R}^\ell$, where ℓ may be large relative to the size of the dataset. Classic results from nonparametric statistics indicate that minimax rates for learning functions conditional on high-dimensional covariates can be poor when the conditional expectation function lacks sufficient smoothness (Stone, 1982). Concerns about high-dimensional covariates, the stringent smoothness requirements they impose, and their impact on efficient semiparametric inference have motivated the development of higher-order influence function (HOIF) estimators (Robins et al., 2008, 2017).

Fortunately, Assumption 3 obviates many of these concerns, including the need to implement HOIF-based estimators: when the annotation score function is known, a first-order influence function-corrected estimator has no (asymptotic) bias, which is what HOIF-based estimators seek to correct (see, e.g., Section 1.1 of Liu et al. (2023) for concise discussion of this). Moreover, Assumption 3 facilitates Assumption 5, which says that we only need to care about the universal consistency of the high-dimensional regression function we must learn, not its rate. Still, when it comes to learning a high-dimensional nonparametric regression function from unstructured data in practice, classic minimax-style results seem too pessimistic, as success in learning high-dimensional nonparametric regressions in finite samples is exactly what has powered much of the progress in the development of AI over the last decade—plausibly due to the benefit of enormous amounts of "pretraining" data. This will be discussed further when considering applications of MARS where Assumption 3 is relaxed.

A natural question related to discussion of high-dimensional nonparametric regression is: why make use of the high-dimensional unstructured data U at all, given that identification of θ holds conditional only on low-dimensional X, per Assumption 2? Intuitively, this is because supplementing X with U in the conditional expectation function μ unlocks the highest quality predictions, which can drastically increase the precision of estimates.[6] Not including U in $\tilde{X}$ is tantamount to trying to summarize the information in structured data M^* (e.g., a text topic) based on U (e.g., the text) without looking at U (the text) at all! Theoretically, including information on U in $\tilde{X}$ can be thought of as enabling even smaller semiparametric variance lower bounds, as assessed from the perspective of comparing efficient estimators across different possible choices of $\tilde{X}$. This result is explored briefly in the following section.

Third, *unstructured datasets can be massive.* When working with massive-scale unstructured datasets, the structured data of interest may represent a "rare event." For example, the structured data may indicate whether a piece of text in a large database (*e.g.,* all social media posts or historical news articles) pertains to a specific topic, which is rare relative to the dataset's enormity. In rare event estimation, a common concern is the large "coefficient of variation," defined as the ratio of the variance of the event indicator to the probability of the event. This

[6]However, by Assumption 2, U drops out of the annotation score function, keeping it low-dimensional, and maintaining strong overlap.

can make uncertainty quantification, such as constructing confidence intervals, less informative.

To address this, as suggested by the rare event estimation literature, one might optimize the annotation function π to reduce the variance of estimates. This could involve incorporating elements of importance sampling (Sigman, 2010). It is worth noting that this notion of efficiency is distinct from the semiparametric efficiency discussed earlier, which treats π as fixed. Recent work by Zrnic and Candes (2024) proposes an importance sampling-inspired solution called "batch active inference," which we revisit in Section 6.1.

Very large dataset size also connects to the concept of "weak overlap" in observational causal inference (D'Amour et al., 2021; Ma et al., 2023) and "decaying overlap" in semi-supervised inference (Zhang et al., 2023). In the MARS framework, we do not assume an asymptotic scenario where $P(A = 1) \to 0$. Instead, we focus on scenarios where a non-negligible fraction of the unstructured dataset can be labeled. Future research could extend the MARS framework to incorporate this alternative asymptotic perspective. For practical applications, we recommend variants of batch active inference for working with massive unstructured datasets.

3.4 Optimal Feature Selection for Imputation

In order to motivate why we may choose to include U in $\tilde{X}$, we must first define a core object of interest in this paper: the MARS mean functional.

Definition 1. We define a "MARS mean functional" as any functional that can be written as

$$\theta(P) = E_P[\tilde{M}^*]$$

where $\tilde{M}^* = g(M^*, V)$ for a known deterministic function g and known random variable V (which is not itself a function of $\pi(X)$) with $[V \perp\!\!\!\perp A] \mid X$, where g is homogeneous of degree one in its first argument.

As will become clear in the following section of this paper, many functionals of missing structured data—including all the functionals of missing structured data considered in this paper—can be written as MARS mean functionals. For such functionals, we have the following identification result.

Proposition 1. *Under Assumptions 1 and 2, a MARS mean functional $E_P[\tilde{M}^*]$ can always be point identified as*

$$E_P[\tilde{\mu}(\tilde{X})],$$

where $\tilde{\mu}(\tilde{X}) := E[\tilde{M} \mid A = 1, \tilde{X}]$, for $\tilde{M} := g(M, V)$ and $\tilde{X} := (X, U, V)$.

Intuitively, then, if we observe some ground truth values of the structured data $\tilde{M}^*$ (i.e., $P(A = 1) > 0$), then we can identify θ. Importantly, so long as $(X, V) \in \tilde{X}$, this identification result holds, which is to say that point identification holds under many possible choices of $\tilde{X}$, so long as any added components to $\tilde{X}$ are still independent of annotation given X. The next two propositions help us unpack what a good choice of $\tilde{X}$ might look like.

To start, we need to compute the efficient influence functional for a MARS mean functional, which we introduce in the next result.

Proposition 2. *The efficient influence function for functional* $E_P\left[\tilde{\mu}(\tilde{X}_i)\right]$*, under the semiparametric statistical model* $\mathcal{P}_\pi \ni P$ *induced by Assumption 3, and under Assumption 2, is*

$$\varphi(W) = \tilde{\mu}(\tilde{X}) + \frac{A}{\pi(X)}(\tilde{M} - \tilde{\mu}(\tilde{X})) - \theta,$$

where π *is the annotation scoring function.*

We know that the semiparametric variance lower bound for θ will be given by $\text{Var}(\varphi)$ for any given statistical model based on iid data $(\tilde{M}, A, \tilde{X})$. In the following proposition, we highlight the dependence of $\text{Var}(\varphi)$ on $\tilde{X}$ itself, which echoes similar statements made by studies of (nonparametric) regression adjustment in experiments (e.g., Cytrynbaum (2021); List et al. (2024)).

Proposition 3. *Under Assumption 2, the semiparametric variance lower bound for functional* $E_P\left[\tilde{\mu}(\tilde{X}_i)\right]$ *is given by*

$$Var(\varphi) \equiv Var(\varphi)(\tilde{X}) = Var(\tilde{M}^*) + E[(\pi(X)^{-1} - 1)\, Var(\tilde{M}^* \mid \tilde{X})].$$

Under a constant annotation score π*,*

$$Var(\varphi)(\tilde{X}) = Var(\tilde{M}^*)\left(1 + \left(\frac{1}{\pi} - 1\right)\left(1 - R^2(\tilde{X})\right)\right),$$

where $R^2(\tilde{X}) := 1 - \frac{E[\text{Var}(\tilde{M}^*|\tilde{X})]}{\text{Var}(\tilde{M}^*)}$*, the nonparametric* R^2*.*

For intuition about the benefit of including U in $\tilde{X}$, the second expression in Proposition 3 is especially instructive. It tell us that choices of $\tilde{X}$ that have higher nonparametric R^2 deflate the semiparametric variance lower bound, with the lowest possible bound achieved for by a $\tilde{X}$ that explains all variation in $\tilde{M}^*$, which recovers $\text{Var}(\varphi) = \text{Var}(\tilde{M}^*)$, the asymptotic variance we would expect if no data were missing. As such, if U dramatically increases nonparametric R-squared, then we want to include it in our $\tilde{X}$. Moreover, we notice that other "comparative statics" for this expression are quite sensible, i.e., if $\pi = 1$, which is to say we annotate all data, then we again recover $\text{Var}(\varphi) = \text{Var}(\tilde{M}^*)$, the asymptotic variance we would expect if no data were missing.

3.5 MARS Algorithm

We now provide an overview of how to implement the MARS framework in a general setting. At a high level, MARS follows a canonical procedure for efficient and robust semiparametric inference, a method with origins in Pfanzagl (1982) that has gained significant attention in biostatistics and econometrics in recent years, particularly for causal inference (e.g., Bang and Robins (2005); van der Laan and Rubin (2006); Chernozhukov et al. (2018)). In this paper, we specifically build on frameworks for deriving efficient and robust semiparametric estimators as discussed in works such as Hines et al. (2022) and Kennedy (2023). MARS adopts this canonical procedure to missing structured data using the following steps:

1. **Identification:** A researcher starts with a target functional $\theta : \mathcal{P} \to \mathbb{R}$ of interest, e.g., the mean of some distribution, a coefficient from a linear regression model, or an average causal

effect. The MARS framework requires that this parameter would be point identified if not for missing structured data. Consistency of potential outcomes and missing at random structured data (assumptions 1 and 2) will then allow the researcher to recover point identification for their target functional under missing structured data.

2. **Deriving the efficient influence function:** If the point identified target functional is pathwise differentiable,[7] then it has a unique efficient influence function (EIF). There are many approaches and algorithms for computing the EIF for a functional. A particularly straightforward approach for doing so is outlined in Kennedy (2023), which we use to construct an EIF candidate for the examples discussed in this paper.

3. **Constructing the robust and efficient estimator:** With the EIF in hand, the researcher may follow one of (at least) three different procedures for forming a robust and efficient estimator: adding a "one-step correction" to a plug-in estimator based on the EIF; solving an "estimating equation" based on the EIF (which aligns most closely with the framework of Chernozhukov et al. (2018), discussed later); or pursuing a targeted maximum likelihood estimation (TMLE) procedure (van der Laan and Rubin, 2006). In this paper, a one-step correction is used, although any of these methods would work interchangeably, offering slightly differing profiles of benefits, e.g., finite sample properties, or ease of derivation.

4. **Sample splitting for estimation:** Regardless of the estimator constructed, the researcher will implement estimation via a data splitting (or cross-fitting) procedure. Although sample splitting is not necessary under certain assumptions about the class of first-step estimators (e.g., Chen et al. (2024)), it is an agnostic way to ensure asymptotically efficient estimation in the large sample limit.

More intuition is provided through the various examples developed in the following section.

4 Applications of MARS

We illustrate the MARS framework in five empirical settings of particular interest to economists: descriptive moments, linear regression, linear instrumental variables (IV) models, difference-in-differences (DiD) designs, and regression discontinuity designs (RDD) under local randomization. In all settings considered, we denote missing structured data as M, and unstructured data as X. We develop each example by assigning a single specific variable to be M (*e.g.*, an outcome or treatment), although MARS can be equally applied to settings where alternative - or multiple - variables are imputed from unstructured data. MARS is limited to pathwise differentiable functionals, in the sense that $\sqrt{n}$-consistent estimators are not guaranteed to exist for nonpathwise differentiable functionals, and as such efficiency would be ill defined. Future work could seek to provide better guidance for consistent and asymptotically normal estimation with unstructured data in settings with nonpathwise differentiable functionals.

We begin by stating the following lemma, which greatly simplifies the derivations of efficient influence functions for the applications considered.

[7] Pathwise differentiability is, intuitively, a smoothness requirement for the functional of interest. For a formal description of pathwise differentiability, see, e.g., van der Vaart (1998), or various proofs in the Appendix, which rest on this definition.

Lemma 1. *The efficient influence function for a point identified MARS mean functional $E_P[\tilde{\mu}(\tilde{X})]$ under a nonparametric statistical model $\mathcal{P} \ni P$ is the same as the efficient influence function of $E_P[\tilde{\mu}(\tilde{X})]$ under the semiparametric statistical model $\mathcal{P}_\pi \ni P$ induced by Assumption 3.*

Lemma 1 is useful for deriving efficient influence functions for mean functionals in the MARS framework. Indeed, all of the applications we consider involve missing data functionals that can be cast as mean functionals. Under the MARS framework, the statistical model of the data under consideration is semiparametric, since we know the annotation scoring function π, per Assumption 3. Deriving efficient influence functions for pathwise differentiable functionals under semiparametric statistical models is typically more challenging than doing so under fully nonparametric statistical models, for which there is only one influence function, which is the efficient influence function. (Under MARS, the statistical model would be nonparametric if π was not known.) Lemma 1 simplifies our task by showing that the EIF for a MARS mean functional θ under a nonparametric statistical model is also the correct EIF for θ under the semiparametric statistical model defined by the MARS assumptions. This implies that we only need to derive the EIF in the fully nonparametric setting, as it remains valid even when incorporating the semiparametric constraints. Intuitively, this lemma holds because perturbing the distribution given by the annotation score of a MARS mean functional does not change the value of the functional. If we had labeled our data in a different (but valid) way, the population-level value of the parameter being estimated would remain unchanged.

4.1 Descriptive Moments

We begin by applying MARS to descriptive moments. In addition to being of relevance to many economic analyses, these functionals will lay the foundation for analyzing more complicated estimands.

Consider the dataset $\{W_i\}_{i=1}^n$ where $W_i := (M_i, A_i, X_i, U_i)$, with variable of interest $M_i \in \mathcal{M} \times \{0\} \subseteq \mathbb{R}$, annotation indicator $A_i \in \{0, 1\}$, observed covariates $X_i \in \mathcal{X} \subseteq \mathbb{R}^k$, and unstructured data $U_i \in \mathcal{U} \subseteq \mathbb{R}^\ell$. We further assume that $W_i \overset{\text{iid}}{\sim} P$, for some arbitrary distribution P. We define $\tilde{X}_i := (X_i, U_i)$.

We wish to compute a descriptive moment of the data, and without loss of generality will focus on the mean.[8] Assuming that $E[|M_i|] < \infty$, we define the estimand of interest, denoted by θ, as the expected value of M_i^*:

$$\theta := E[M_i^*].$$

We use the terms estimand, functional, and parameter interchangeably to refer to θ.

Under Assumptions 1 (consistency of potential outcomes) and 2 (missing at random structured data), we can identify θ as

$$\theta = E\left[\mu(\tilde{X}_i)\right],$$

where $\mu(\tilde{X}_i) := E[M_i \mid A_i = 1, \tilde{X}_i] = E[M_i \mid A_i = 1, X_i, U_i]$.

[8] This is without loss of generality, insofar as M may be defined as some underlying structured random variable raised to a particular power.

Proposition 4. *The efficient influence function for functional* $E_P\left[\mu(\tilde{X}_i)\right]$*, under the semi-parametric statistical model* $\mathcal{P}_\pi \ni P$ *induced by Assumption 3, and under Assumption 2, is*

$$\varphi(W_i) = \mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i)) - \theta,$$

where π *is the annotation scoring function.*

As such, we can form the (doubly) robust and efficient one-step estimator as

$$\hat{\theta} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i))\right],$$

where $\mathcal{I}$ is the set of indices of the data allocated to the "estimation" partition of a random data split, and where $\hat{\mu}$ (the deep neural network) is estimated on the other "training" partition, represented by the set of indices $\mathcal{I}'$, i.e., $\hat{\mu}$ is a random function of $\{W_{i'}\}_{i'\in\mathcal{I}'}$. Note that this is the exactly the augmented inverse probability weighted (AIPW) estimator, which arises as the efficient estimator of a potential mean in causal inference settings (Robins et al., 1994; Rotnitzky et al., 1998; Scharfstein et al., 1999). The reason for this is clear: M^* is just a potential outcome, as discussed in Section 3.1.

The following theorem confirms that this estimator is indeed efficient, and robust in the sense that the only assumption we need to make about first-step (or, nuisance) estimation is the very mild Assumption 5.

Theorem 1. *Under Assumptions 1 (consistency of potential outcomes), 2 (missing at random structured data), 3 (known and bounded annotation score function) and 5 (MSE consistency of the imputation function), we have that*

$$\sqrt{|\mathcal{I}|}(\hat{\theta} - \theta) \underset{d}{\to} N(0,\, Var(\varphi(W_i))).$$

Remark 1 (AIPW, prediction-powered inference, and flexible regression adjustment)**.** *We now show that, for mean estimation, a statistical literature on prediction-powered inference (PPI) and recent work in economics on flexible regression adjustment (FRA) can both be viewed as special cases unified under the banner of efficient and robust semiparametric inference for missing data, the perspective taken by MARS.*

Suppose that $\pi(X_i) = |\mathcal{I}\cap\mathcal{J}|/|\mathcal{I}|$*, where* $\mathcal{J}$ *is the set of indices corresponding to annotated data points. Then we see that*

$$\begin{aligned}
\hat{\theta} &= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i))\right] && \text{(AIPW)}\\
&= \underbrace{\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\hat{\mu}(\tilde{X}_i)}_{A} + \underbrace{\frac{1}{|\mathcal{I}\cap\mathcal{J}|}\sum_{i\in\mathcal{I}\cap\mathcal{J}}(M_i^* - \hat{\mu}(\tilde{X}_i))}_{B} && \text{(PPI)}\\
&= \underbrace{\frac{1}{|\mathcal{I}\cap\mathcal{J}|}\sum_{i\in\mathcal{I}\cap\mathcal{J}} M_i^*}_{C} + \underbrace{\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\hat{\mu}(\tilde{X}_i) - \frac{1}{|\mathcal{I}\cap\mathcal{J}|}\sum_{i\in\mathcal{I}\cap\mathcal{J}}\hat{\mu}(\tilde{X}_i)\right)}_{D}. && \text{(FRA)}
\end{aligned}$$

The term A in the second expression (PPI) is the best imputation-based guess of $E[M_i^*]$ *in*

the estimation sample (it is the naive "plug-in" estimator), and the term B is a bias correction term—it can be thought of as an estimate of the measurement error of our imputation function in the annotated sample. This expression is reproduced in recent work on "prediction-powered inference" (Angelopoulos et al., 2023). As noted in Angelopoulos et al. (2023), in settings where $\mathcal{I}$ is very large—which is common in unstructured data applications—the variance of term A will be small. If the researcher ignores the fact that the structured data are imputed, they would report very precise estimates. The variance of term B will depend on the predictive performance of the imputation function $\hat{\mu}$, as well as the size of the annotated set $\mathcal{J}$. The better the imputation function, and the larger the $|\mathcal{J}|$, the shorter the confidence intervals. If our imputation function is not very predictive, running it over a massive dataset will contain limited information about the underlying true distribution of data, which term B accounts for.

The last expression (FRA) is reported in recent work on "flexible regression adjustment" (List et al., 2024). Under this formulation, we can view term C as our best estimate of the quantity of interest using only ground truth data. We then leverage the imputation function $\hat{\mu}$ as a form of nonparametric regression adjustment in term D, with the same intuition as a linear regression adjustment: we adjust for systematic differences between our large unlabeled sample and our small annotated ground truth sample as summarized by $\hat{\mu}$.

4.2 Linear Regression

We now apply MARS to a model that is ubiquitous in the social sciences: linear regression. In particular, we consider the case where an outcome variable is measured from unstructured data, though the MARS framework can readily handle if one or more regressors (as well as the outcome) are instead based on unstructured data.

Consider the dataset $\{W_i\}_{i=1}^n$, where $W_i = (M_i, C_i, A_i, X_i, U_i)$, with outcome $M_i \in \mathcal{M} \times \{0\} \subseteq \mathbb{R}$, regressors $C_i \in \mathcal{C} \subseteq \mathbb{R}^d$, annotation indicator $A_i \in \{0, 1\}$, observed covariates $X_i \in \mathcal{X} \subseteq \mathbb{R}^k$, and unstructured data $U_i \in \mathcal{U} \subseteq \mathbb{R}^\ell$. We further assume that $W_i \overset{\text{iid}}{\sim} P$ for some distribution P, and that $[C_i \perp\!\!\!\perp A_i] \mid X_i$.

We assume that

$$M_i^* = C_i^{\mathsf{T}}\theta + \varepsilon_i, \quad E[\varepsilon_i \mid C_i] = 0,$$

where $\theta \in \Theta \subseteq \mathbb{R}^p$. We are interested in identifying and estimating the j-th regression coefficient θ_j.

By the Frisch-Waugh-Lovell theorem, we have that

$$\theta_j = E\left[{C_{i,j}^{\perp}}^2\right]^{-1} E\left[C_{i,j}^{\perp} M_i^*\right],$$

where $C_{i,j}^{\perp} := C_{i,j} - E^*[C_{i,j} \mid 1, C_{i,1}, \ldots, C_{i,j-1}, C_{i,j+1}, \ldots, C_{i,d}]$, and where E^* is the linear projection operator.

Let $\tilde{X}_i := (X_i, U_i, C_i)$. Then, by Assumptions 1 (consistency of potential outcomes) and 2 (missing at random structured data), we can identify θ_j as

$$\theta_j := \theta_{j,\text{den}}^{-1}\theta_{j,\text{num}} = E\left[{C_{i,j}^{\perp}}^2\right]^{-1} E\left[C_{i,j}^{\perp}\mu(\tilde{X}_i)\right]$$

where $\mu(\tilde{X}_i) = E\left[M_i \mid A_i = 1, \tilde{X}_i\right]$.

Proposition 5. *The efficient influence function for the numerator functional* $E_P\left[C_{i,j}^{\perp}\mu(\tilde{X}_i)\right]$ *under the semiparametric statistical model* $\mathcal{P}_\pi \ni P$ *induced by Assumption 3, and under Assumption 2, is*

$$\varphi_{num}(W_i) = C_{i,j}^{\perp}\left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i))\right] - \theta_{j,num},$$

where π *is the annotation scoring function. The efficient influence function for the denominator functional* $E_P[{C_{i,j}^{\perp}}^2]$ *is given by*

$$\varphi_{den}(W_i) = {C_{i,j}^{\perp}}^2 - \theta_{j,den}.$$

We can then form the robust and efficient one-step estimators for both the numerator and denominator functionals as

$$\hat{\theta}_{j,\text{num}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\check{C}_{i,j}\left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i))\right], \quad \hat{\theta}_{j,\text{den}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\check{C}_{i,j}^2,$$

where $\check{C}_{i,j} := C_{i,j} - \hat{E}^*[C_{i,j} \mid 1, C_{i,1}, \ldots, C_{i,j-1}, C_{i,j+1}, \ldots, C_{i,d}]$, and where $\hat{E}^*$ is the least squares operator.

Combining these estimators, we can form the efficient estimator for θ_j as

$$\hat{\theta}_j := \hat{\theta}_{j,\text{num}}\hat{\theta}_{j,\text{den}}^{-1} = \frac{\sum_{i\in\mathcal{I}}\check{C}_{i,j}\left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i))\right]}{\sum_{i\in\mathcal{I}}\check{C}_{i,j}^2}.$$

In particular, note that this is just a residualized regression with pseudo-outcome $\hat{\varphi}_i := \hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i))$, i.e.,[9]

$$\hat{\theta}_j := \hat{\theta}_{j,\text{num}}\hat{\theta}_{j,\text{den}}^{-1} = \frac{\sum_{i\in\mathcal{I}}\check{C}_{i,j}\hat{\varphi}_i}{\sum_{i\in\mathcal{I}}\check{C}_{i,j}^2}.$$

Theorem 2. *Under Assumptions 1 (consistency of potential outcomes), 2 (missing at random structured data), 3 (known and bounded annotation score function) and 5 (MSE consistency of the imputation function), we have that*

$$\sqrt{|\mathcal{I}|}(\hat{\theta}_j - \theta_j) \underset{d}{\to} N(0, \mathit{Var}(\varphi_j(W_i)))$$

where

$$\varphi_j(W_i) := \theta_{j,den}^{-1}(\varphi_{num}(W_i) - \theta_j\varphi_{den}(W_i))$$

is the efficient influence function for θ_j*.*

Importantly, we notice in this application that the imputation function $\hat{\mu}$ is a function of context specific variables as well, i.e., $\tilde{X}_i := (X_i, U_i, C_i)$. In the machine learning literature, learning a function of both unstructured and structured (tabular) data is sometimes called "wide and deep learning" (Cheng et al., 2016), and has been further studied in recent works in econometrics (Klaassen et al., 2024). Note that this is because, under Assumption 2, the

[9]The use of the term "pseudo-outcome" here mirrors similar language used by Egami et al. (2023) for debiased inference, and echoes terminology from elsewhere in causal inference.

imputation function should asymptotically approximate the function $E[M \mid A = 1, X = x, U = u, C = c]$ in order for estimation to be efficient, absent further assumptions.

In many settings, a researcher may perform annotation such that $[C_j^{\perp} M \perp\!\!\!\perp A] \mid X$ also holds, in which case an "optimal" imputation function could approximate $E[C_j^{\perp} M \mid A = 1, X = x, U = u]$ instead. However, this function would be more difficult to learn in practice, requiring that the researcher build a predictor of $C_j^{\perp} M$ solely from unstructured data features U and observed covariates relevant to labeling X, a plausibly fraught learning problem. Why? Though U (and to a lesser extent X) very plausibly carries signal about M^*, e.g., unstructured text data U is useful for predicting the text topic M^*, neither U nor X may carry any obvious signal about $C_j^{\perp}$, e.g., a control regressor that is in no way a summary of the text, nor how it co-varies with M^*. By augmenting our unstructured data with C from the start, we side-step such issues entirely.

If we additionally made the assumption that $[C_j^{\perp} \perp\!\!\!\perp M^*] \mid (X, U)$, then the optimal imputation function would no longer need to be a function of C, i.e., optimal imputation now attempts to approximate $E[M \mid A = 1, X = x, U = u] = E[M^* \mid X = x, U = u]$, a "direct" imputation function. This sort of assumption is, however, not without loss, and may be false, whereas previous constructions based on including C remain agnostic about the (conditional) independence of $C_j^{\perp}$ and M^*. As will be discussed later on, even if this new conditional independence assumption did not hold, imputation functions that ignored variation in C would still yield *valid* inference under Assumption 3, but would not yield (asymptotically) *efficient* inference. In other words, imputation functions that ignore dependence on C, without further assumption, are non-optimal (in the sense of the efficiency of the estimator they produce).

4.3 Linear Instrumental Variables

The MARS framework is straightforward to extend to linear IV. In this section, we follow the terminology and setup of Blandhol et al. (2022).

Consider dataset $\{W_i\}_{i=1}^n$ with

$$W_i = (Y_i, M_i, C_i, Z_i, A_i, X_i, U_i),$$

where $Z_i \in \mathcal{Z} \subseteq \mathbb{R}$ is a candidate instrumental variable (IV), $C_i \in \mathcal{C} \subseteq \mathbb{R}^d$ are covariates (which contain a constant), $M_i \in \mathcal{M} = \{0, 1\}$ is a treatment of interest, $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ is an outcome of interest, $A_i \in \{0, 1\}$ is the annotation indicator, $X_i \in \mathcal{X} \subseteq \mathbb{R}^k$ are observed covariates relevant to annotation, and $U_i \in \mathcal{U} \subseteq \mathbb{R}^{\ell}$ are unstructured data. We further assume that $W_i \overset{\text{iid}}{\sim} P$ for some joint distribution P, and that $[(C_i, Z_i) \perp\!\!\!\perp A_i] \mid X_i$.

Potential outcomes are denoted $\{Y_i^m\}_{m \in \mathcal{M}}$ and potential treatments are denoted $\{M_i^z\}_{z \in \mathcal{Z}}$. Assuming consistency of potential outcomes and treatment, we have

$$Y_i = \sum_{m \in \mathcal{M}} \mathbb{I}(M_i = m) Y_i^m, \quad M_i = \sum_{z \in \mathcal{Z}} \mathbb{I}(Z_i = z) M_i^z.$$

We are interested in identifying and estimating the average treatment effect:

$$\theta := E[Y_i^{m=1} - Y_i^{m=0}].$$

Per Blandhol et al. (2022), we assume that the candidate instrumental variable is both relevant and exogenous, that treatment effects are the same for all individuals, and that the conditional expectation of the potential outcome is a linear function of the covariates. We can then express θ as

$$\theta = E[Z_i^{\perp} M_i^*]^{-1} E[Z_i^{\perp} Y_i],$$

where $Z_i^{\perp}$ is Z_i linearly residualized with C_i, using the same notation as in the previous application. Under these assumptions, θ is exactly the two stage least squares (TSLS) estimand.[10]

Let $\tilde{X}_i := (X_i, U_i, Z_i, C_i)$. Then, by Assumptions 1 (consistency of potential outcomes) and 2 (missing at random structured data), we can identify θ as

$$\theta = \theta_{\text{den}}^{-1} \theta_{\text{num}} = E[Z_i^{\perp} \mu(\tilde{X}_i)]^{-1} E[Z_i^{\perp} Y_i]$$

where $\mu(\tilde{X}_i) := E[M_i \mid A_i = 1, \tilde{X}_i]$.

Proposition 6. *The efficient influence function for the numerator functional $E_P\left[Z_i^{\perp} Y_i\right]$ under the semiparametric statistical model $\mathcal{P}_\pi \ni P$ induced by Assumption 3 is*

$$\varphi_{num}(W_i) = Z_i^{\perp} Y_i - \theta_{num},$$

and the efficient influence function for the denominator functional $E_P[Z_i^{\perp} \mu(\tilde{X}_i)]$ under $\mathcal{P}_\pi \ni P$, and under Assumption 2, is

$$\varphi_{den}(W_i) = Z_i^{\perp} \left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)} (M_i - \mu(\tilde{X}_i)) \right] - \theta_{den}.$$

As with the linear regression example, we can form the efficient estimator for θ by first forming efficient one-step estimators for θ_{num} and θ_{den},

$$\hat{\theta}_{\text{num}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \check{Z}_i Y_i, \quad \hat{\theta}_{\text{den}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \check{Z}_i \left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i, Z_i, C_i)} (M_i - \hat{\mu}(\tilde{X}_i)) \right],$$

and then combining them to form

$$\hat{\theta} = \hat{\theta}_{\text{num}} \hat{\theta}_{\text{den}}^{-1} = \frac{\sum_{i \in \mathcal{I}} \check{Z}_i Y_i}{\sum_{i \in \mathcal{I}} \check{Z}_i \left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)} (M_i - \hat{\mu}(\tilde{X}_i)) \right]},$$

where $\check{Z}_i$ is Z_i residualized with a least squares fit on C_i.

Note that the efficient estimator is just a TSLS regression with pseudo-treatment $\hat{\varphi}_i := \hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)} (M_i - \hat{\mu}(\tilde{X}_i))$, or

$$\hat{\theta} = \hat{\theta}_{\text{num}} \hat{\theta}_{\text{den}}^{-1} = \frac{\sum_{i \in \mathcal{I}} \check{Z}_i Y_i}{\sum_{i \in \mathcal{I}} \check{Z}_i \hat{\varphi}_i}.$$

Theorem 3. *Under Assumptions 1 (consistency of potential outcomes), 2 (missing at random structured data), 3 (known and bounded annotation score function) and 5 (MSE consistency of*

[10] If we were to drop Assumptions CLE and LIN, per Blandhol et al. (2022), and instead impose their "ordered strong monotonicity" (OSM) assumption, and if C_i were "rich," e.g., C_i were saturating indicators of the relevant *discrete* covariate space for (conditional) exogeneity to hold, then θ would still have a "weakly causal" interpretation.

the imputation function), we have that

$$\sqrt{|\mathcal{I}|}(\hat{\theta} - \theta) \underset{d}{\to} N(0, \, Var(\varphi(W_i)))$$

where

$$\varphi(W_i) = \theta_{den}^{-1}(\varphi_{num}(W_i) - \theta\varphi_{den}(W_i))$$

is efficient influence function for θ*.*

Notice once again that the optimal imputation function—which facilitates efficient estimation without further assumption—is also a function of Z and C.

4.4 Differences-in-differences

In this application, we focus on the nonparametrically founded differences-in-differences (DiD) estimator introduced in Callaway and Sant'Anna (2021).

Consider the dataset $\{W_i\}_{i=1}^n$:

$$W_i = (M_{i1}, \ldots M_{iT}, D_{i1}, \ldots, D_{iT}, A_{i1}, \ldots, A_{iT}, X_{i1}, \ldots, X_{iT}, U_{i1}, \ldots, U_{iT}),$$

where $M_{it} \in \mathcal{M} \times \{0\} \subseteq \mathbb{R}$ is an outcome of interest, $D_{it} \in \mathcal{D} = \{0, 1\}$ are treatment indicators, $A_{it} \in \{0, 1\}$ are annotation indicators, $X_{it} \in \mathcal{X} \subseteq \mathbb{R}^k$ are observed covariates, and $U_{it} \in \mathcal{U} \subseteq \mathbb{R}^\ell$ are unstructured data. Suppose that $W_i \overset{\text{iid}}{\sim} P$ for some joint distribution P, and that $[D_{it} \perp\!\!\!\perp A_{it'}] \mid X_i$ for all t, t'.

Let G_{ig} be a binary indicator for a unit i first being treated at time g and let C_i be an indicator for units that are never treated. Furthermore, we define $M_{it}^{g=0}$ to be the untreated potential outcome of a unit i at time t and M_{it}^g to be the potential outcome of a unit i at time t if they first became treated in period g. Assuming consistency of potential outcomes, we have

$$M_{it} = M_{it}^{g=0} + \sum_{g=2}^{T} \left(M_{it}^g - M_{it}^{g=0} \right) G_{ig}.$$

We are interested in the estimand

$$\theta = E\left[M_{it}^{*\,g} - M_{it}^{*\,g=0} \mid G_{ig} = 1 \right],$$

the average treatment effect on those treated in cohort g at time t. This is the crucial input to any other substantive DiD estimand of interest, which will just be a known or identified weighted average of these estimands. For concreteness, but without loss of generality, we set $g = 2$ and $t = 2$, the canonical two-period DiD setup. Under Assumptions 1–4 of Callaway and Sant'Anna (2021), assuming that Assumptions 3 and 4 hold unconditionally, as well as under the non-anticipation assumption of Callaway and Sant'Anna (2021), we can express θ as

$$\theta = E\left[M_i^* \mid G_{i2} = 1 \right] - E\left[M_i^* \mid C_i = 1 \right]$$

for $M_i^* := M_{i2}^* - M_{i1}^*$.

Let $M_i := M_{i2} - M_{i1}$, $A_i := A_{i1}A_{i2}$, and $\tilde{X}_i := (X_{i1}, X_{i2}, U_{i1}, U_{i2})$. Then under Assumptions 1 (consistency of potential outcomes) and 2 (missing at random structured data), we can identify

θ as

$$\theta = \theta_G - \theta_C, \quad \theta_G = E[\mu_G(\tilde{X}_i) \mid G_{i2} = 1], \quad \theta_C = E[\mu_C(\tilde{X}_i) \mid C_i = 1],$$

where $\mu_G(\tilde{X}_i) = E[M_i \mid G_{i2} = 1, A_i = 1, \tilde{X}_i]$ and $\mu_C(\tilde{X}_i) = E[M_i \mid C_i = 1, A_i = 1, \tilde{X}_i]$.

Proposition 7. *The efficient influence functions for functionals $E_P[\mu_G(X_i) \mid G_{i2} = 1]$ and $E_P[\mu_C(X_i) \mid C_i = 1]$ under the semiparametric statistical model $\mathcal{P}_\pi \ni P$ induced by Assumption 3, and under Assumption 2, are, respectively,*

$$\varphi_G(W_i) = \frac{G_{i2}}{P(G_{i2} = 1)} \left[\mu_G(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu_G(\tilde{X}_i)) \right] - \theta_G,$$
$$\varphi_C(W_i) = \frac{C_i}{P(C_i = 1)} \left[\mu_C(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu_C(\tilde{X}_i)) \right] - \theta_C.$$

The efficient influence function for $E_P[\mu_G(X_i) \mid G_{i2} = 1] - E_P[\mu_C(X_i) \mid C_i = 1]$ under $\mathcal{P}_\pi \ni P$ is then given by

$$\varphi(W_i) = \varphi_G(W_i) - \varphi_C(W_i).$$

We can form efficient one-step estimators for each θ_G and θ_C as

$$\hat{\theta}_G = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{G_{i2}}{\hat{P}(G_{i2} = 1)} \left[\hat{\mu}_G(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}_G(\tilde{X}_i)) \right],$$
$$\hat{\theta}_C = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{C_i}{\hat{P}(C_i = 1)} \left[\hat{\mu}_C(\tilde{X}_i) - \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}_C(\tilde{X}_i)) \right],$$

where $\hat{P}$ is an empirical probability, and we can combine them to form the efficient estimator for θ as

$$\hat{\theta} = \hat{\theta}_G - \hat{\theta}_C.$$

Theorem 4. *Under Assumptions 1 (consistency of potential outcomes), 2 (missing at random structured data), 3 (known and bounded annotation score function) and 5 (MSE consistency of the imputation function), we have that*

$$\sqrt{|\mathcal{I}|}(\hat{\theta} - \theta) \underset{d}{\to} N(0, \mathit{Var}(\varphi(W_i)))$$

where $\varphi(W_i)$ is the efficient influence function for θ.

4.5 RDD under Local Randomization

Lastly, we consider a regression discontinuity design (RDD) under MARS. We focus on sharp RDD under the local randomization framework, as opposed to the continuity framework (Cattaneo and Titiunik, 2022).

Consider the dataset $\{W_i\}_{i=1}^n$ with $W_i = (M_i, R_i, D_i, A_i, X_i, U_i)$, where $M_i \in \mathcal{M} \times \{0\} \subseteq \mathbb{R}$ is an outcome of interest, $R_i \in \mathcal{R} \subseteq \mathbb{R}$ is the running variable, $D_i \in \mathcal{D} = \{0, 1\}$ is a treatment indicator, $A_i \in \{0, 1\}$ is the annotation indicator, $X_i \in \mathcal{X} \subseteq \mathbb{R}^k$ are observed covariates, and $U_i \in \mathcal{U} \subseteq \mathbb{R}^\ell$ are unstructured data. We assume $W_i \overset{\text{iid}}{\sim} P$ for some joint distribution P, and that

$[(R_i, D_i) \perp\!\!\!\perp A_i] \mid X_i$.

Potential outcomes $\{M_i^d\}_{d\in\mathcal{D}}$ are related to observed outcomes via the assumption of consistency:

$$M_i = D_i M_i^{d=1} + (1 - D_i)M_i^{d=0}.$$

We are interested in the estimand

$$\theta = E\left[M_i^{*d=1} - M_i^{*d=0} \mid R_i \in \mathcal{B}\right]$$

for a set (or "window") $\mathcal{B} \subset \mathcal{R}$. Under Assumptions LR1 and LR2 of Cattaneo et al. (2024), which enforce that local randomization holds in $\mathcal{B}$, we can write θ as

$$\theta = E[M_i^* \mid R_i \in \mathcal{B}, D_i = 1] - E[M_i^* \mid R_i \in \mathcal{B}, D_i = 0].$$

Let $\tilde{X}_i := (X_i, U_i)$. Then, under Assumptions 1 (consistency of potential outcomes) and 2 (missing at random structured data), we can identify θ as $\theta = \theta_1 - \theta_0$ where

$$\theta_d = E[\mu_d(\tilde{X}_i) \mid R_i \in \mathcal{B}, D_i = d],$$

with $\mu_d(\tilde{X}) = E[M_i \mid R_i \in \mathcal{B}, D_i = d, A_i = 1, \tilde{X}_i]$.

Proposition 8. *The efficient influence function for $E_P[\mu_d(\tilde{X}_i) \mid R_i \in \mathcal{B}, D_i = d]$ under the semiparametric statistical model $\mathcal{P}_\pi \ni P$ induced by Assumption 3, and under Assumption 2, is*

$$\varphi_d(W_i) = \frac{\mathbb{I}\{R_i \in \mathcal{B}, D_i = d\}}{P(R_i \in \mathcal{B}, D_i = d)}\left[\mu_d(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu_d(\tilde{X}_i))\right] - \theta_d.$$

As such, the efficient influence function for functional $E_P[\mu_1(\tilde{X}_i) \mid R_i \in \mathcal{B}, D_i = 1] - E_P[\mu_0(\tilde{X}_i) \mid R_i \in \mathcal{B}, D_i = 0]$ under $\mathcal{P}_\pi$ and Assumption 2 is

$$\varphi(W_i) = \varphi_1(W_i) - \varphi_0(W_i).$$

Forming efficient one-step estimators for θ_1 and θ_0, respectively, as

$$\hat{\theta}_1 = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \frac{\mathbb{I}\{R_i \in \mathcal{B}, D_i = 1\}}{\hat{P}(R_i \in \mathcal{B}, D_i = 1)}\left[\hat{\mu}_1(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}_1(\tilde{X}_i))\right],$$
$$\hat{\theta}_0 = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \frac{\mathbb{I}\{R_i \in \mathcal{B}, D_i = 0\}}{\hat{P}(R_i \in \mathcal{B}, D_i = 0)}\left[\hat{\mu}_0(\tilde{X}_i) - \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}_0(\tilde{X}_i))\right],$$

where $\hat{P}$ again indicates an empirical probability, we can combine them to form an efficient estimator of θ as

$$\hat{\theta} = \hat{\theta}_1 - \hat{\theta}_0,$$

just as we did for the DiD application considered earlier.

Theorem 5. *Under Assumptions 1 (consistency of potential outcomes), 2 (missing at random structured data), 3 (known and bounded annotation score function) and 5 (MSE consistency of*

the imputation function), we have that

$$\sqrt{|\mathcal{I}|}(\hat{\theta} - \theta) \underset{d}{\to} N(0, \mathit{Var}(\varphi(W_i)))$$

where $\varphi(W_i)$ is the efficient influence function for θ.

Remark 2 (Pathwise differentiability and RDD under the continuity framework)**.** *Under the continuity framework, the analogous sharp RDD estimand involves functionals of the form*

$$\theta(r) := E[M_i^* \mid R_i = r],$$

where $R_i \in \mathcal{R}$ is presumed to be continuous, and $\theta(r)$ is to be evaluated at a particular (limiting) point in $\mathcal{R}$. In the literature on semiparametric inference, it is well known that functionals of this type (expectations conditional on continuous random variables evaluated at a specific point) are not pathwise differentiable, and hence there is no regular, $\sqrt{n}$-consistent estimator of the quantity, let alone an efficient estimator in the sense previously discussed (Hines et al., 2022). That said, frameworks for valid asymptotic inference on such functionals (that incorporate flexible first-step estimators) do exist (Bibaut and van der Laan, 2017; Noack et al., 2024), and resemble strategies arising from inference on conditional average treatment effects with continuous conditioning covariates in the causal inference literature (Kennedy et al., 2017; Semenova and Chernozhukov, 2021).

5 MARS and Other Frameworks

5.1 MARS and Frameworks for Inference using Predictions from Black-Box Machine Learners

Recent work under the title of "prediction-powered inference" (Angelopoulos et al., 2023, 2024; Zrnic and Candès, 2024; Zrnic and Candes, 2024; Ji et al., 2025; Kluger et al., 2025) and "design-based supervised learning" (Egami et al., 2023, 2024) shows that, given access to a validation sample, under the appropriate debiased construction of an estimator, a researcher can perform valid large sample inference leveraging predictions of variables of interest from arbitrary, black-box machine learning models f. These predictions may be thought of exactly as imputations based on arbitrary imputation function f. MARS extends these frameworks, making clear exactly what the machine learning model f should be (or should approximate) in order to make such estimators *efficient*, i.e., MARS sheds light on the *optimal* f.

Unpacking this literature in the context of MARS, note that even if the imputation function is not universally consistent for μ, we can still perform valid, unbiased large sample inference if Assumption 3 (known and bounded annotation score function) holds, though this inference is no longer guaranteed to be efficient. Formally, consider that, instead of learning a function $\hat{\mu}$ which is universally consistent for μ, we opt to use the external, fixed machine learning model f as our imputation function. This f could be, for example, a pretrained, black-box deep neural network—such as a LLM—that is capable of reasonably good zero-shot imputation. For exposition, consider applying f as an imputation function in the simplest application of MARS,

descriptive moments. Now consider the estimator

$$\hat{\theta}_f = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left[f(\tilde{X}_i) + \frac{A_i}{\pi(X_i)} (M_i - f(\tilde{X}_i)) \right].$$

By Assumption 3 (known and bounded annotation score function), even if f is meaningfully different from μ (and will not converge to μ, given that f is fixed as the sample size grows in the asymptotic thought experiment), it is nearly immediate by the construction of this estimator that $\hat{\theta}_f$ is consistent and unbiased for θ. The fixed imputation function-based estimator $\hat{\theta}_f$ is also just an average of iid random variables, and so the Lindeberg-Lévy CLT applies here, with

$$\sqrt{|\mathcal{I}|}(\hat{\theta}_f - \theta) \underset{d}{\to} N(0, \sigma^2)$$

where $\sigma^2 := \text{Var}\left(f(\tilde{X}_i) + \frac{A_i}{\pi(X_i)} (M_i - f(\tilde{X}_i)) \right)$. This is the sort of valid—but possibly inefficient—inference that previous frameworks in this literature guarantee. Double robustness is seen in the fact that knowledge of the annotation (propensity) score permits the researcher to use even highly misspecified imputation functions f without spoiling the validity of inference (just its variance).

MARS, through its analysis of semiparametric efficiency, demonstrates that even the most accurate "direct" black-box imputation methods, i.e., $f(x) = E[M^* \mid \tilde{X} = \tilde{x}]$ exactly, can be non-optimal in certain common settings, and should incorporate context-specific variables beyond unstructured data to produce the most precise estimates. This, to our knowledge, has not been emphasized in the existing literature.

5.2 MARS and Double/Debiased Machine Learning

Many of the MARS estimators resemble those found in the literature on semiparametric inference with missing data in biostatistics and econometrics. As a notable example, descriptive moment estimation under missing structured data produces an estimator that is identical to the augmented inverse probability weighted (AIPW) estimator (Robins et al., 1994). This estimator is also equivalent to one derived via the Neyman orthogonal score for estimating a potential mean under the "selection on observables" assumption in a causal inference setting, as outlined in the double/debiased machine learning (DML) framework (Chernozhukov et al., 2018).

This equivalence is not coincidental. In Section 3.1, we demonstrate that the expected value of missing structured data can be interpreted as an average potential outcome, and point identification is achievable under what amounts to an "annotation on observables" assumption. Furthermore, deriving a Neyman orthogonal score can be viewed as an "estimating equations" approach to constructing semiparametrically efficient estimators (Kennedy, 2023; Schuler and van der Laan, 2024), as opposed to alternative methods such as the one-step correction (explored in MARS) or the targeted maximum likelihood estimation (TMLE) framework of van der Laan and Rubin (2006).

Because MARS is based on the same fundamental semiparametric analysis as DML, there may be many ways to cross-pollinate insights from work done for the DML framework with MARS. Consider a recent strand of the DML literature that has focused on "automatic" or data-driven ways of implementing debiasing corrections (Chernozhukov et al., 2022a,b). Although

many functionals considered under MARS lead to Riesz representers with simple and common analytic expressions, there are other missing structured data settings that may induce more complicated functionals, and automatic debiasing may prove quite useful in these settings.

6 Extensions to MARS

We now consider extensions to MARS that are useful when the unstructured datasets under consideration are very large, *e.g.,* millions to billions of instances.

6.1 Batch Active Inference (Zrnic and Candes, 2024)

Unlike in the observational causal inference setting where the propensity score function is unknown, in the unstructured data setting we plausibly get to choose the annotation score function, π. "Batch active inference" Zrnic and Candes (2024) provides one useful approach for choosing this function in the very large database setting.

For intuition, consider that, when working with very large unstructured datasets, it will often be the case that missing structured data of interest constitutes a "rare event." For example, if the dataset is the universe of all social media posts on a major platform, the share of content about almost any given topic of interest will be extremely small, given the vast diversity of content. When it comes to rare event estimation, the typical concern is that the "coefficient of variation" (i.e., the ratio of the variance of the indicator of the event over the probability of the event) is quite large, making confidence intervals uninformative. As the literature on rare event estimation would suggest, we may be interested in annotation that incorporates some element of importance sampling in order to further reduce the variance of our estimates (Sigman, 2010).

In the setting of descriptive moment estimation, recall that the asymptotic variance of the estimator $\hat{\theta}$ is given by

$$\begin{aligned}\sigma^2 &:= \text{Var}\left(\mu(\tilde{X}) + \frac{A}{\pi(X)}(M - \mu(\tilde{X}))\right) \\ &= E\left[\frac{A}{\pi(X)^2}(M - \mu(\tilde{X}))^2\right] + c\end{aligned}$$

where c is a constant that collects all terms that, under our assumptions, do not vary with $\pi(X)$. To find the importance sampling-based annotation score function, we then solve

$$\min_{\pi,\lambda,\kappa}\left\{E\left[\frac{A}{\pi(X)^2}(M - \mu(\tilde{X}))^2\right] + \lambda(E[\pi(X)] - 1) - \kappa E[\pi(X)]\right\}$$

where (λ, κ) are Karush-Kuhn-Tucker (KKT) multipliers that enforce $0 \leq E[\pi(X)] \leq 1$. In light of Assumption 2 (missing at random structured data), the optimal π_{IS} is given by

$$\pi_{IS}(X) \propto E\left[(M^* - \mu(\tilde{X}))^2 \mid X\right]^{1/2}.$$

This annotation function is infeasible, as it depends on M^*, which we do not observe prior to annotation, but it provides a useful intuition: the annotation score function should place more weight on sampling data points that are harder to impute, in the sense of MSE. However, as advocated by Zrnic and Candes (2024), we can consider implementing a feasible annotation

score function that captures this intuition, such as

$$\pi_{FIS}(X) \propto \text{err}(X) \approx E\left[(M^* - \mu(\tilde{X}))^2 \mid X\right]^{1/2}$$

where $\text{err}(x)$ is some proxy function that captures the uncertainty of a fixed imputation function. For example, distributional characteristics of the softmax scores of deep neural encoder outputs or verbalized confidence scores from LLMs may serve as such a proxies (Yang et al., 2024).

6.2 Functionals of Aggregated Missing Structured Data

When working with a large unstructured database, it is often the case that the functional of interest concerns the distribution of (a transformation of) *aggregated* missing structured data, and the researcher does not observe any ground truth at this level of aggregation.

Formally, consider the dataset $\{W_{ij}\}_{i\in[N_j],j\in[n]}$, $W_{ij} \overset{\text{iid}}{\sim} P_1$, $N_j \overset{\text{iid}}{\sim} P_2$, where i indexes the granular, "record" level at which the unstructured data lives, and j indexes some level of aggregation. For example, i may index individual newspaper articles, and j may index counties in which those articles are run. Let $W_{ij} = (M_{ij}, A_{ij}, X_{ij}, U_{ij})$, with observed variable of interest $M_{ij} \in \mathcal{M} \times \{0\} \subseteq \mathbb{R}$, annotation indicator $A_{ij} \in \{0, 1\}$, observed covariates $X_{ij} \in \mathcal{X} \subseteq \mathbb{R}^k$, and unstructured data $U_{ij} \in \mathcal{U} \subseteq \mathbb{R}^\ell$. We suppose that Assumptions 1 and 2 hold at the unstructured data "record" level i:

$$M_{ij} = A_{ij}M^*_{ij}, \quad [M_{ij} \perp\!\!\!\perp A_{ij}] \mid X_{ij}.$$

Suppose that the functional of interest is a descriptive mean for the distribution of structured data aggregated at level j:

$$\theta = E[M^*_j], \quad M^*_j = \alpha(\{M^*_{ij}\}_{i=1}^{N_j}),$$

where α is an operator that returns a scalar quantity. Unfortunately, this missing data functional is generically *not* point identified, insofar as we do not observe any ground truth for aggregated M^*_j, i.e., $A_j = 0$ with probability 1.

However, in many settings of interest, the aggregation of interest will be some known linear combination of the record-level structured data, under which we may recover point identification. To represent this setting, we introduce the following assumption.

Assumption 6 (Linearly aggregated structured data)**.** *For record-level structured data $M^*_{ij} \overset{iid}{\sim}$ and record count $N_j \overset{iid}{\sim}$, we define linearly aggregated structured data as*

$$M^*_j := \omega_j \sum_{i=1}^{N_j} M^*_{ij}.$$

We further assume that $N_j \perp\!\!\!\perp M^*_{ij}$ and $\omega_j = h(N_j)$ for some measurable function h.

If Assumption 6 holds, by the law of total expectation we have that

$$\begin{aligned}
\theta &= E[M_j^*] \\
&= E\left[\omega_j \sum_{i=1}^{N_j} M_{ij}^*\right] \\
&= E\left[E\left[\omega_j \sum_{i=1}^{N_j} M_{ij}^* \mid N_j\right]\right] \\
&= E\left[\omega_j N_j E\left[M_{ij}^* \mid N_j\right]\right] \\
&= E\left[\omega_j N_j E\left[M_{ij}^*\right]\right] \\
&= E\left[\omega_j N_j\right] E\left[M_{ij}^*\right],
\end{aligned}$$

allowing us to write

$$\theta = \theta_1 \theta_2, \quad \theta_1 := E\left[\omega_j N_j\right], \quad \theta_2 := E\left[M_{ij}^*\right].$$

This intuitive decomposition of θ can be interpreted as the product of the expected (effective) aggregate group size θ_1 and the expected record-level mean θ_2. Importantly, θ_2 is point identified, because we observe ground truth at the record level. Further note that if $\omega_j = 1$, i.e., we are interested in sums, then $\theta = E[N_j]E[M_{ij}^*]$. If $\omega_j = 1/N_j$, i.e., we are interested in averages, then simply $\theta = E[M_{ij}^*]$.

Then, so long as Assumptions 3 and 5 also hold at granular record level i, we know using the exact same set of results from Theorem 1 that an efficient estimator for θ_2 is

$$\hat{\theta}_2 = \frac{1}{|\mathcal{I}|} \sum_{j,i \in \mathcal{I}} \left[\hat{\mu}(\tilde{X}_{ij}) + \frac{A_{ij}}{\pi(X_{ij})}(M_{ij} - \hat{\mu}(\tilde{X}_{ij}))\right]$$

where $\hat{\mu}(\tilde{X}_{ij}) := \hat{E}[M_{ij} \mid A_{ij} = 1, \tilde{X}_{ij}]$. We can also simply and efficiently estimate θ_1 with a plug-in method, such as

$$\hat{\theta}_1 := \frac{1}{n} \sum_{j=1}^{n} \omega_j N_j.$$

These two estimates can be combined to produce an efficient estimator $\hat{\theta}$ for θ, using identical logic to that of the estimators in the linear regression and linear IV examples.[11]

In certain other settings, the aggregation of interest may be a nonlinear transformation of a linear aggregation of record-level structured data. We introduce the following assumption to represent this case.

Assumption 7 (Nonlinearly transformed linearly aggregated structured data)**.** *For record-level structured data $M_{ij}^* \overset{iid}{\sim}$ and record count $N_j \overset{iid}{\sim}$, we define nonlinearly transformed linearly aggregated structured data as*

$$\tilde{M}_j^* := f(M_j^*), \quad M_j^* := \omega_j \sum_{i=1}^{N_j} M_{ij}^*.$$

[11] Note that N_j must be the data on the relevant total record count, i.e., if a researcher is interested in aggregating indicators for all newspaper articles in a county, N_j must be a count of *all* newspaper articles published in a county, not just the total of articles for that county in the sample.

We further assume that $N_j \perp\!\!\!\perp M^*_{ij}$, $\omega_j = h(N_j)$ for some measurable function h, and f is twice differentiable.

A relevant example is a researcher interested in inference on the logarithm of some level j averaged structured data,

$$\theta = E[\tilde{M}^*_j] = E\left[\log\left(\frac{1}{N_j}\sum_{i=1}^{N_j} M^*_{ij}\right)\right].$$

In this setting, without ground truth at the aggregate level j, we do not recover point identification of the functional of interest. However, given sufficient smoothness of the transforming function, we can identify what may be a good approximation of θ. Specifically, a Taylor expansion around $E[M^*_j]$ shows that

$$\begin{aligned}
\theta &= E\left[f\left(M^*_j\right)\right] \\
&= f\left(E\left[M^*_j\right]\right) + \frac{1}{2}f''\left(E\left[M^*_j\right]\right)\mathrm{Var}\left(M^*_j\right) + E[R(M^*_j)] \\
&= f\left(\theta_1\theta_2\right) + \frac{1}{2}f''\left(\theta_1\theta_2\right)\theta_3 + E[R(M^*_j)] \\
&= \underline{\theta} + E[R(M^*_j)],
\end{aligned}$$

where

$$\underline{\theta} := f\left(\theta_1\theta_2\right) + \frac{1}{2}f''\left(\theta_1\theta_2\right)\theta_3,$$

$E[R(M^*_j)]$ is a remainder term, and

$$\theta_3 := \mathrm{Var}(\omega_j N_j)E[M^*_{ij}]^2 + E[\omega_j^2 N_j]\mathrm{Var}(M^*_{ij}),$$

which is identified with ground truth at record level i.[12] As such, we can point identify $\underline{\theta}$, because, intuitively, we have linearized our functional of interest θ.

When is $\underline{\theta}$ close to θ? In other words, when is this second order Taylor expansion a good approximation, or when is $E[R(M^*_j)]$ sufficiently small? Note that

$$E[R(M^*_j)] = E\left[\frac{1}{6}f'''\left(\xi\left(M^*_j\right)\right)\left(M^*_j - E[M^*_j]\right)^3\right]$$

for $\xi(M^*_j)$ which lies between M^*_j and $E[M^*_j]$ with probability 1, by the mean value representation of the Taylor remainder. We anticipate this term to be small, for example, when $\left|f'''\left(\xi\left(M^*_j\right)\right)\right| < C$ with probability 1 (such as when M^*_j is bounded with probability 1), and that M^*_j has a law similar to a normal random variable (which we may indeed think is the case

[12] Notice, for example, when the aggregation of interest is an average, θ_3 simplifies as

$$\theta_3 = E[M^*_{ij}]^2 + E[N_j^{-1}]\mathrm{Var}(M^*_{ij}).$$

for averages, such as $\omega_j = 1/N_j$ and N_j large), i.e.,

$$\begin{aligned} E[R(M_j^*)] &= E\left[\frac{1}{6} f'''\left(\xi\left(M_j^*\right)\right)\left(M_j^* - E[M_j^*]\right)^3\right] \\ &\leq \frac{C}{6} E\left[\left(M_j^* - E[M_j^*]\right)^3\right] && \text{(Boundedness.)} \\ &\approx 0. && \text{(Small third central moment.)} \end{aligned}$$

6.3 Linear Models with MARS First-Step Measurement Error

One common use case for the MARS framework—as will be explored in the following section—involves using MARS to estimate a set of (crystallized) random variables that then act as regressors in a downstream linear regression model. This setting follows closely to estimation of coefficients in linear models with regressors estimated from randomly sampled survey data; see, e.g., Deaton (1985) for notable work in economics, which itself is based on seminal work in measurement error models compiled by Fuller (1987).

To start, consider the dataset $\{W_i\}_{i=1}^n$, where $W_i \overset{\text{iid}}{\sim} P$, and $W_i = (X_i^*, Y_i)$, with $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ an outcome of interest and $X_i \in \mathcal{X} \subseteq \mathbb{R}^K$ regressors. Without loss of generality, $E[W_i] = 0$ (both variables are centered). We consider the linear model

$$Y_i = X_i^{*\mathsf{T}}\beta + \varepsilon_i, \quad E[\varepsilon_i \mid X_i^*] = 0.$$

Dropping subscripts as a stacked matrix-vector notation, we can write

$$Y = X^*\beta + \varepsilon,$$

where X^* is a $n \times K$ matrix, and β, Y, and ε are each $K \times 1$ vectors.

In particular, we consider the case that X_i^* is not actually observed, but has been estimated with a MARS first-step, X_i. Accordingly, then, we know that

$$X_{k,i} = X_{k,i}^* + \eta_{k,i}, \quad (\varepsilon_i, X_{k,i}^*) \perp\!\!\!\perp \eta_{k,i}, \quad \eta_{k,i} \overset{d}{\approx} N(0, \sigma_{\eta,k}^2), \quad \sigma_{\eta,k}^2 := |\mathcal{I}_k|^{-1}\text{Var}(\varphi_k),$$

where φ_k is the efficient influence function associated with X_k. As such, we are in a classical measurement error setting. This setup nests the case where only some regressors are generated by a MARS first-step, in which case $\sigma_{\eta,k}^2 = 0$ for non-MARS regressors.

We therefore will assume that

$$\eta_i \perp\!\!\!\perp Y_i, \quad \eta_i \overset{\text{iid}}{\sim} N(0, \Sigma), \quad \Sigma := \text{diag}(\sigma_{\eta,1}^2, \ldots, \sigma_{\eta,K}^2),$$

owing to the facts that (a) it will often be the case that each MARS first-step X_k will be estimated on a separate sample, and (b) for large $\mathcal{I}_k$, the normal approximation should be quite accurate. We consider that Σ is known, a not unreasonable assumption for sufficiently large $\mathcal{I}_k$.

Under this model, note that β satisfies the GMM moment condition

$$E\left[g(X_i, Y_i, \beta)\right] = 0, \quad g(X_i, Y_i, \beta) = X_i(Y_i - X_i^{\mathsf{T}}\beta) + \Sigma\beta.$$

Correspondingly, the consistent, plug-in GMM estimator is given by

$$\hat{\beta} := \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^{\mathsf{T}} - \Sigma\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) = (X^{\mathsf{T}}X - n\Sigma)^{-1}X^{\mathsf{T}}Y,$$

and possesses the usual just identified GMM asymptotic distribution, under additional mild regularity conditions,

$$\sqrt{n}(\hat{\beta} - \beta) \underset{d}{\to} N(0, G^{-1}\Xi G^{-1})$$

where $G = E[\nabla_\beta g(X_i, Y_i, \beta)] = -(E[X_i X_i^{\mathsf{T}}] - \Sigma)$, and $\Xi = \mathrm{Var}(g(X_i, Y_i, \beta)) = \mathrm{Var}(X_i(Y_i - X_i^{\mathsf{T}}\beta))$.

Using Isserlis' theorem, we can also re-write the asymptotic variance as

$$\mathrm{AVar}(\hat{\beta}) = \Omega^{-1}E\left[\varepsilon_i^2 X_i^* X_i^{*\mathsf{T}}\right]\Omega^{-1} + \Omega^{-1}\Pi\Omega^{-1}$$

where

$$\Omega = E[X_i^* X_i^{*\mathsf{T}}], \quad \Pi := E[\varepsilon_i^2]\Sigma + \beta^{\mathsf{T}}\Sigma\beta\,(\Omega + \Sigma) + \Sigma\beta\beta^{\mathsf{T}}\Sigma,$$

which is the usual sandwich asymptotic variance of the OLS estimator without measurement error, plus an inflation factor which becomes smaller the more precise our MARS estimates are. (In particular, note that $\Pi = 0$ if $\Sigma = 0$, as we would expect.)

We can form consistent plug-in estimators for all of these quantities, such that

$$\hat{V}^{-1/2}\sqrt{n}(\hat{\beta} - \beta) \underset{d}{\to} N(0, I)$$

where

$$\hat{V} = \hat{G}^{-1}\hat{\Xi}\hat{G}^{-1} = (\hat{E}[X_i X_i^{\mathsf{T}}] - \Sigma)^{-1}\widehat{\mathrm{Var}}(X_i(Y_i - X_i^{\mathsf{T}}\hat{\beta}))(\hat{E}[X_i X_i^{\mathsf{T}}] - \Sigma)^{-1}.$$

We are thus able to produce asymptotically valid confidence intervals with desired coverage.

Note that this setup can be straightforwardly extended to the case of clustering, using the appropriate clustered CLT. It can also be readily specialized to the panel data case, which was the original setting of interest to Deaton (1985). Moreover, there are various arms reach extensions of the framework for cases where, e.g., there is heterogeneity in variance for η_i across i (see Section 9.3); the outcome is a MARS first-step as well (or alone); measurement error distributions are not assumed to be normal; that Σ must be estimated; and so on.[13]

The MARS first-step is crucial for making this linear measurement error model plausible, in particular providing a credible reason to believe in classical measurement error, and, moreover, showing that returns to prediction accuracy in a MARS first-step will trickle down to increased precision of a linear regression model coefficient.

7 Empirical Examples

This section illustrates MARS with three empirical examples. The first two examples apply MARS to prominent papers in the economics literature that make use of unstructured data—Baker et al. (2016) and Caldara and Iacoviello (2022)—while the final example probes MARS

[13] Also see, e.g., Theorems 2.2.1 and 3.1.1 of Fuller (1987) for examples of statements of asymptotic distributions under such enriched models.

design choices using a mean estimation example that builds upon annotated data previously collected by the authors (Dell et al., 2023).

7.1 The Economic Policy Uncertainty Index (Baker et al., 2016)

Baker et al. (2016) is an influential recent paper in empirical macroeconomics that introduces a quantitative index to measure economic policy uncertainty (EPU). The EPU index calculates at each point in time the proportion of articles published in leading newspapers that discuss EPU, which is measured using a simple, keyword-based binary classifier.[14] These proportions are then scaled and normalized to produce the final index values.

Intuitively, then, each entry in the EPU index can be viewed as an estimate of the probability that a newspaper article reflects economic policy uncertainty—in the sense defined by Baker et al. (2016)—at a specific point in time for a given set of newspapers. This is an inference problem where unstructured data (newspaper text) are used to estimate a functional of missing structured data (the expectation of a binary indicator for economic policy uncertainty at a given point in time). This makes the estimation of the EPU index a suitable setting for applying the MARS framework.

For fixed time period t, consider the dataset $\{W_{it}\}_{i=1}^{n}$, where $W_{it} = (M_{it}, A_{it}, U_{it})$. Here, $M_{it} \in \{0,1\}$ is a binary random variable indicating the expression of economic policy uncertainty by newspaper article i at time t, $A_{it} \in \{0,1\}$ is an annotation indicator for article i at time t, and $U_{it} \in \mathcal{U} \subseteq \mathbb{R}^{\ell}$ is newspaper article text for article i at time t. We assume $W_{it} \overset{\text{iid}}{\sim} P_t$, that is, the newspaper articles are an iid sample from some superpopulation of articles of interest, period by period. Moreover, in this empirical setting we know that $(U_{it}, M_{it}^*) \perp\!\!\!\perp A_{it}$ and that the annotation score π is constant (described shortly), and so Assumption 2 is guaranteed to hold.

We may represent the EPU index's value at time t as

$$\text{EPU}_t := c_t \theta_t, \quad \theta_t := E_{P_t}[M_{it}^*],$$

where c_t is a constant specific to time t that incorporates known information on overall economic uncertainty at time t, as well as any other normalizing constants of interest. Written this way, it is clear that θ_t is a descriptive moment functional of the form discussed in Section 4.1, and as such we know that, under Assumptions 1 (consistency of potential outcomes), 2 (missing at random structured data), 3 (known and bounded annotation score function), and 5 (MSE consistency of the imputation function), the robust and semiparametric efficient estimator of θ_t is given by

$$\hat{\theta}_t = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left[\hat{\mu}(U_{it}) + \frac{A_{it}}{\pi} (M_{it} - \hat{\mu}(U_{it})) \right],$$

where $\hat{\mu}(U_{it}) := \hat{E}[M_{it} \mid A_{it} = 1, U_{it}]$.

We can therefore compute the "MARS-EPU index" as $\{c_t \hat{\theta}_t\}_{t \in \mathcal{T}}$, for any given $\mathcal{T} \subset \mathbb{N}$, a set of time period indices. In particular, we compute a yearly MARS-EPU index where $\mathcal{T} = \{1985, 1986, \ldots, 2010\}$ (which Baker et al. (2016) term the modern—in contrast to the historical—EPU index) using two different imputation functions: the same keyword classifier implemented for the original modern EPU index, and a deep neural classifier based on AllenAI's

[14] For details on the keyword classifier training algorithm, see Section III.A of Baker et al. (2016).

Longformer model (Beltagy et al., 2020), which is well-suited for encoding texts of newspaper article length. To remain faithful to the definition of economic policy uncertainty implemented by Baker et al. (2016), we leverage only the ground truth annotations that these authors created for their human audit sample—that is, we do not label any additional data ourselves—and additionally randomly sample unannotated articles for each time period from ProQuest at a ratio of 9 unannotated articles for every annotated article. Given that all data are iid randomly sampled, the annotation score function for our final sample is $\pi = 0.1$.[15] We employ a simple data-splitting technique, in which half the labeled data (across all periods) is used to train and validate the imputation function, which is deployed for all $t \in \mathcal{T}$, and the other half is used to estimate the EPU index for all $t \in \mathcal{T}$. Cross-fitting could be used to further improve precision, although care must be taken in how training and validation data for the classifier are partitioned from estimation data (see, e.g., Bach et al. (2024) for discussion of related issues in the context of hyperparameter tuning).

The Longformer model deployed is small in parameter count (it is a "base"-sized model with 149M parameters) and was trained on a single Nvidia RTX A6000 GPU card (comparable in quality to a GPU available cheaply through a cloud computing provider) to obtain an accuracy of 81.2% in the estimation sample. The Baker et al. (2016) keyword classifier, by contrast, achieves an accuracy of 65.3% in the estimation sample.

In Figure 1, we plot both modern MARS-EPU indices along with counterpart modern EPU indices in the fashion of the original paper, which we write as $\{c_t\tilde{\theta}_t\}_{t\in\mathcal{T}}$, where

$$\tilde{\theta}_t = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \hat{\mu}(U_{it}).$$

We compute 95% confidence intervals for all index values using consistent, plug-in estimators of the asymptotic variances of the estimators based on the known form of the efficient influence function.[16]

Most notably, Figure 1 illustrates that the confidence intervals for the EPU index should be much larger if one properly accounts for uncertainty due to correction for systematic measurement error. Indeed, the unadjusted EPU index confidence intervals are only generically valid if one is interested in inference on $E[\hat{\mu}(U_{it})]$ for arbitrary $\hat{\mu}$, as opposed to $E[M^*_{it}]$—or if one assumes that $\hat{\mu}$ is an unbiased estimator of μ, which is seemingly rejected by apparent, though mild, downward bias in many years of the index—and are otherwise anti-conservative.

Because the annotation score function is known, the confidence intervals reported for both versions of the MARS-EPU index are valid asymptotically. However, the better the imputation function is—the closer it is to the conditional expectation function μ—the narrower the confidence intervals will be for a given n. Hence, the MARS-EPU index based on a deep neural classifier has shorter confidence intervals than the MARS-EPU index based on the Baker et al.

[15] As in the original paper, all data, both annotated and unannotated, is randomly sampled *conditional* on the text in question meeting certain preliminary keyword requirements, namely the inclusion of specific keywords pertaining to the economy and to uncertainty. As such, the proper interpretation of the estimand θ_t is one that construes the superpopulation of interest as leading newspaper articles that contain specific economics and uncertainty keywords. The constants c_t are then primarily responsible for modifying the interpretation of the EPU index, effectively changing the denominator of the index values to be reflective of the full population of newspaper articles of interest, under the key assumption that no articles discussing EPU are devoid of the aforementioned economics and uncertainty keywords used for filtering.

[16] See Appendix Section 9.2 for a formal statement and proof.

(2016) keyword classifier. This reflects that the returns to predictive accuracy will manifest in confidence interval length, a point similarly made by, e.g., Angelopoulos et al. (2023). In our experience, a frequent question from those new to deep learning is how accurate a classifier needs to be: the answer depends on the required precision, as evinced by Figure 1.

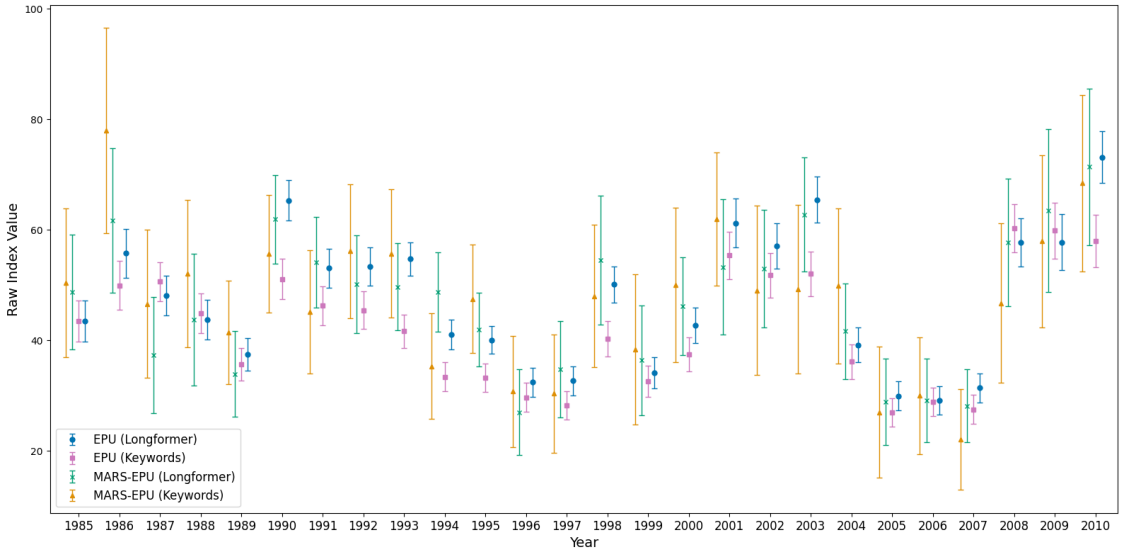


Figure 1: Modern EPU and MARS-EPU Indices (1985-2010).

In Baker et al. (2016), much of the social scientific value of the EPU index is realized by using the index as a regressor in downstream regression analyses. We now revisit a representative regression analysis from the paper, and investigate to what extent application of the MARS framework to the EPU index changes its conclusions.

In particular, we reanalyze the following baseline regression reported in Table IV, column (5) of Baker et al. (2016):

$$\Delta \text{Emp}_{it} = \beta \Delta \log \text{EPU}_t \times \text{intensity}_{it} + \gamma \Delta \frac{\text{Federal purchases}_t}{\text{GDP}_t} \times \text{intensity}_{it} + \alpha_i + \psi_t + u_{it},$$

where i indexes firms and t indexes years; α_i and ψ_t are firm and year fixed effects, respectively; intensity_{it} is a firm-year policy exposure intensity variable, ΔEmp_{it} is a firm-year employment growth variable, and $\Delta \frac{\text{Federal purchases}_t}{\text{GDP}_t}$ is a relevant control variable; and u_{it} is an error term. The estimand of interest is β.

In this setting, we recognize that the regressor for the coefficient of interest must be estimated, and, if done so using a MARS first-step, we may apply the framework introduced in Section 6.3 to recover valid inference on β (or, more specifically, the extended framework for independent but not identically distributed measurement error outlined in Appendix Section 9.3). Note that, per our MARS first-step, we have that

$$\sqrt{|\mathcal{I}_t|}\left(\hat{\theta}_t - \theta_t\right) \underset{d}{\to} N(0, \text{Var}(\varphi_t)),$$

and so, recalling that $\text{EPU}_t := c_t \theta_t$,

$$\sqrt{|\mathcal{I}_t|}\left(\widehat{\text{EPU}}_t - \text{EPU}_t\right) \underset{d}{\to} N(0, c_t^2 \text{Var}(\varphi_t)).$$

Using the delta method, then,

$$\sqrt{|\mathcal{I}_t|}\left(\log \widehat{\mathrm{EPU}}_t - \log \mathrm{EPU}_t\right) \underset{d}{\to} N(0, \theta_t^{-2}\mathrm{Var}(\varphi_t)).$$

Motivated by this CLT, we assume that

$$\log \widehat{\mathrm{EPU}}_t = \log \mathrm{EPU}_t + \tilde{\eta}_t, \quad \tilde{\eta}_t \overset{\text{ind}}{\sim} N(0, \sigma_t^2),$$

where $\sigma_t^2 := |\mathcal{I}_t|^{-1}\theta_t^{-2}\mathrm{Var}(\varphi_t)$. Then, by the independence of MARS measurement errors, we also have that

$$\Delta \log \widehat{\mathrm{EPU}}_t = \Delta \log \mathrm{EPU}_t + \check{\eta}_t, \quad \check{\eta}_t \overset{\text{ind}}{\sim} N(0, \sigma_t^2 + \sigma_{t-1}^2).$$

Lastly, we may write

$$\Delta \log \widehat{\mathrm{EPU}}_t \times \mathrm{intensity}_{it} = \Delta \log \mathrm{EPU}_t \times \mathrm{intensity}_{it} + \eta_t,$$

where $\eta_t := \check{\eta}_t \times \mathrm{intensity}_{it}$, mimicking the notation introduced in Section 6.3.

In a regression framework with random regressors, we may observe that the MARS first-step measurement error for an interacted MARS-first step regressor, η_t, is no longer strongly classical, in that η_t is no longer independent of $\Delta \log \mathrm{EPU}_t \times \mathrm{intensity}_{it}$, because they both contain $\mathrm{intensity}_{it}$. Fortunately, the measurement error in this setting featuring an interaction is still *weakly* classical (Schennach, 2022), in that

$$\begin{aligned}
&E[\eta_t \mid \Delta \log \mathrm{EPU}_t \times \mathrm{intensity}_{it}] \\
=&E[\check{\eta}_t \times \mathrm{intensity}_{it} \mid \Delta \log \mathrm{EPU}_t \times \mathrm{intensity}_{it}] \\
=&E[E[\check{\eta}_t \times \mathrm{intensity}_{it} \mid \Delta \log \mathrm{EPU}_t \times \mathrm{intensity}_{it}, \Delta \log \mathrm{EPU}_t, \mathrm{intensity}_{it}] \mid \Delta \log \mathrm{EPU}_t \times \mathrm{intensity}_{it}] \\
=&E[\mathrm{intensity}_{it} E[\check{\eta}_t \mid \Delta \log \mathrm{EPU}_t, \mathrm{intensity}_{it}] \mid \Delta \log \mathrm{EPU}_t \times \mathrm{intensity}_{it}] \\
=&0.
\end{aligned}$$

Accordingly, identification of β still holds, though we now must be careful to note that

$$\begin{aligned}
\Sigma_{it} &:= \mathrm{Var}(\check{\eta}_t \times \mathrm{intensity}_{it}) \\
&= E[\mathrm{intensity}_{it}^2]\mathrm{Var}(\check{\eta}_t) \\
&= E[\mathrm{intensity}_{it}^2](\sigma_t^2 - \sigma_{t-1}^2).
\end{aligned}$$

We assume that $E[\mathrm{intensity}_{it}^2]$ varies over time, and that $E[\mathrm{intensity}_{it}^2]$ for any given t is effectively known, given the size of the panel.

As such, our final estimator is given by

$$\hat{\beta} = (X^{\mathrm{T}}X - n\bar{\Sigma})^{-1}X^{\mathrm{T}}Y,$$

where X is a (residualized) design matrix and $\bar{\Sigma}$ is the averaged Σ_{it}.

In Figure 2, we estimate β using the MARS-EPU index along with the appropriate measurement error corrected least squares estimator $\hat{\beta}$ (ME-LS) and plot confidence intervals based on standard errors clustered at the firm level, as per Baker et al. (2016); we also compute point estimates and confidence intervals for β based on the inappropriate OLS estimator, both when

the MARS-EPU index is used as a first-step and when the unadjusted EPU index is used a first-step.[17]

Importantly, in this plot we notice that:

1. When measurement error is taken into account, we indeed see that the more precisely estimated MARS-EPU index based on the deep neural network imputation function yields a narrower confidence interval on the coefficient of interest.
2. Implementing the measurement-error corrected (ME-LS) estimator matters, and the bias of failing to take measurement error into account is evident when comparing the MARS-EPU index based regression coefficients across the ME-LS and OLS estimators.
3. It so happens in this empirical setting that the point estimates produced from using the MARS-EPU indices with the ME-LS estimator (the estimation strategy advocated for in this paper) yields very similar point estimates to those one would obtain by simply using the unadjusted EPU indices in a standard OLS regression. This is perhaps unsurprising, given that bias from imputation was mild, as was observed in Figure 1. However, we see from Figure 2 that failing to apply the MARS framework would understate uncertainty, even in this case; indeed, applying the MARS framework appropriately shows that a 95% confidence interval should contain zero for the estimated coefficient of interest, unlike the completely unadjusted results.

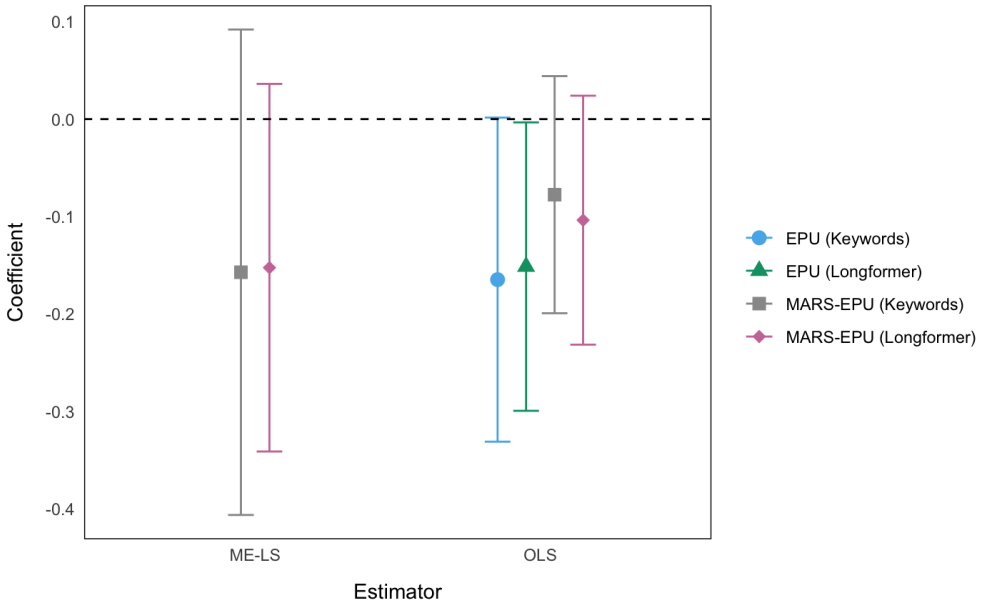


Figure 2: Estimates of β by estimator for EPU and MARS-EPU indices.

[17]Note that the coefficient estimate based on the unadjusted, keyword-imputation-based EPU index and the OLS estimator produced in Figure 2 differs somewhat from the conceptually comparable estimate reported in Baker et al. (2016). This is because the regression results reported in this table are: (1) based on a sample of data from 1985-2010, as opposed to 1985-2012, as in the original paper; and (2) the EPU index deployed in this paper is based on deploying the keyword classification algorithm of Baker et al. (2016) in the validation sample collected for this paper, and as such the final keywords in the classifier differ slightly from those of the original EPU index.

7.2 The Geopolitical Risk Index (Caldara and Iacoviello, 2022)

An intellectual successor to Baker et al. (2016), Caldara and Iacoviello (2022) construct a quantitative index measuring geopolitical risk (GPR). Similar to the EPU index, the GPR index value at each point in time is calculated as the (normalized) share of articles from prominent newspapers that discuss "rising" geopolitical risks, based on a detailed keyword-based text query.[18]

Fundamentally then, again, the GPR index value estimates the probability that a newspaper article discusses rising geopolitical risks, suitably defined, at a specific point in time for a given set of newspapers. This inference problem likewise involves using unstructured data (newspaper text) to estimate a functional of missing structured data (the expected value of a binary indicator for articles discussing geopolitical risk). As such, this is another context where applying the MARS framework is appropriate.

We follow an identical formal setup to the one described in the previous section and produce modern ($\mathcal{T} = \{1985, 1986, \ldots, 2015\}$) GPR indices with and without application of the MARS framework. We again form multiple indices based on two distinct imputation functions: the original keyword query for GPR, and a deep neural classifier based on Sentence-BERT's smallest "general purpose" DistilRoBERTa model (Reimers and Gurevych, 2019).[19] As with the EPU index, we remain faithful to the definition of geopolitical risk crafted by Caldara and Iacoviello (2022) by leveraging only the ground truth annotations that were created for their human audit sample, and randomly sample fresh, unannotated articles for each time period from ProQuest at a ratio of 9 unannotated articles for every annotated article. Given that all data is iid randomly sampled, the annotation score function for our final sample is again $\pi = 0.1$.[20] We employ an identical data splitting procedure as for the MARS-EPU indices. In the estimation sample, the deep neural classifier achieves an accuracy of 83.7%, and the keyword query achieves an accuracy of 70.2%, using the human audit annotations as ground truth.

In Figure 3, we plot both versions of the MARS-GPR and unadjusted GPR indices, along with confidence intervals, constructed as they were for the EPU example. We see that the uncorrected GPR index based on the original keyword query is highly and systematically biased, which we conclude by comparison to its debiased counterparts leveraging the ground truth audit sample through MARS. Evidently, many articles in the audit sample discuss geopolitical risk but do not contain the requisite query-specific keywords, leading the keyword-based classification strategy to universally underestimate geopolitical risk. Again, we see that confidence intervals are (erroneously) much narrower for the non-MARS indices than for the MARS indices. Moreover, once again, the more accurate deep neural classifier generates somewhat narrower year by year confidence intervals than the less accurate keyword classifier, as we would expect, and similarly to the EPU example.

We now revisit a representative regression analysis from Caldara and Iacoviello (2022) that makes use of the GPR index as a regressor. In particular, Caldara and Iacoviello (2022) specify

[18] The query is Boolean and can be interpreted as a binary keyword-based classifier. However, unlike the EPU index, there is no explicit training algorithm: Caldara and Iacoviello (2022) constructed the query using data external to the analytic sample and did not update it based on the analytic sample. For the full query and details of its construction, see Section A.3 of the online appendix in Caldara and Iacoviello (2022).

[19] Specifically, we use the model `all-distilroberta-v1`.

[20] As for the EPU index, audit sampling is conditional on first meeting certain keyword inclusion requirements; this influences the interpretation of downstream estimands and estimates, and motivates correction by a known period-by-period constant, under certain assumptions. See appendix B.6 of Caldara and Iacoviello (2022) for more details.

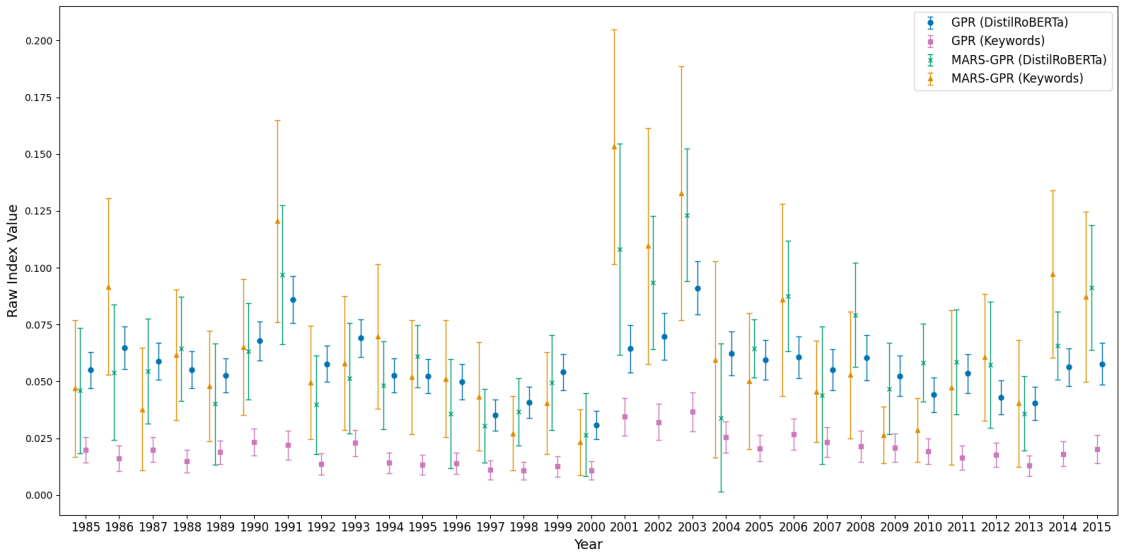


Figure 3: GPR and MARS-GPR Indices (1985-2015).

a simple baseline regression for estimating the impact of GPR on economic disaster probability as

$$D_{it} = \beta \text{GPR}_t + \delta \Delta \text{GDP}_{it-1} + u_{it},$$

where i indexes countries and t indexes years, D_{it} is an economic disaster indicator variable, ΔGDP_{it-1} is a relevant control, and u_{it} is an error term. Again, the estimand of interest is β.

In Figure 4, we similarly estimate β using the MARS-GPR index first-step and the appropriate measurement-error corrected least squares estimator (ME-LS), as well as using a naive OLS estimator with both MARS-GPR and unadjusted GPR first-steps.[21]

In Figure 4, we may notice that:

1. Relative to the ME-LS estimates, the OLS estimates based on the MARS-GPR indices are erroneously precise.
2. Using either of the MARS-GPR indices in the regression analysis yields quantitatively and qualitatively different point estimates of β than those based on either of the uncorrected GPR indices, i.e., non-negligible bias is introduced in estimating β when failing to apply the MARS framework. Moreover, the (lack of) variation in the unadjusted GPR indices leads to relatively much larger confidence intervals for estimates of β.
3. There is effectively no downstream gain in precision from using deep neural imputations in this setting; this too may be unsurprising given the relative performance of the two

[21]Note that the estimate for the regression coefficient β reported in Caldara and Iacoviello (2022) again somewhat differs in magnitude from the comparable estimate produced in Figure 4. This is because the regression results reported in this figure are: (1) based on a sample dating from 1985-2015, as opposed to 1900-2019, as in the original paper; and (2) we do not normalize the GPR index in any way, whereas the results reported in the original paper convert to a scale in units of the standard deviation of the index. We also observe that clustering by both country and year in our restricted 1985-2015 sample, as was performed in the original paper, generates negative variance estimates, even for the OLS estimator. This is known to sometimes occur in multi-way clustering settings (Cameron et al., 2011), and suggests that the asymptotic theory motivating multi-way clustering is a poor fit for this particular application. Dissecting this issue is beyond the scope of our paper, however, and as such we opt to adjust for one-way clustering on unit (country), as was done for the similar application reanalyzed in Baker et al. (2016).

imputation methods observed in the estimation sample.

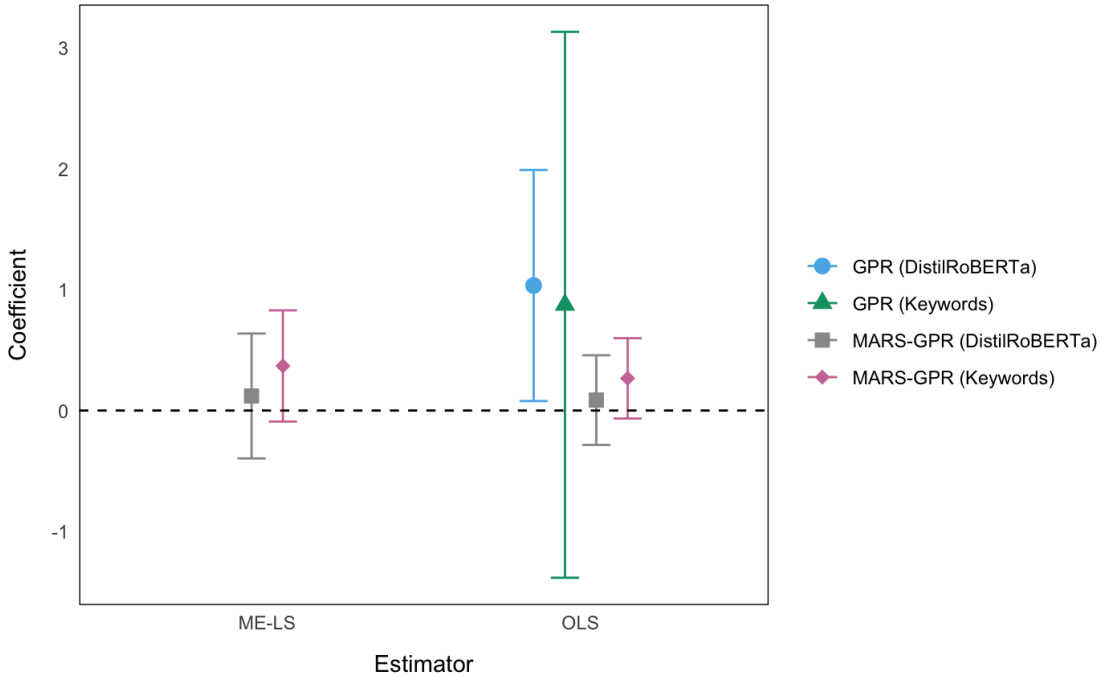


Figure 4: Estimates of β by estimator for GPR and MARS-GPR indices.

7.3 Newspaper Topic Classification (Dell et al., 2023)

In order to probe design choices associated with the MARS framework - in the now familiar context of mean estimation - we develop a third example in which we examine the share of articles that are about politics in local U.S. newspapers across time. Articles were selected at random for annotation from a large sample of historical U.S. newspapers, as described in more detail in Dell et al. (2023). No keyword filters were implemented beforehand, so the relevant population is all front page content in thousands of local newspapers. In the economics literature, annotated text audit samples are typically very small, whereas by labeling our own data, we could create a large enough set to examine how the size of the annotation set influences inference under MARS.

This empirical setting, much as the ones previously explored, can be cast as descriptive moment estimation under the MARS framework, where $M_i \in \{0, 1\}$ is now a binary indicator of whether politics is discussed in the newspaper article. Under Assumptions 1, 2, 3, and 5, we once again have that the semiparametric efficient estimator of the probability of an article being written on the topic of politics is given by $\hat{\theta}$ in section 4.1.

Figure 5 plots estimates $\hat{\theta}$ and associated confidence intervals based on annotating increasingly large fractions of the sample. All estimates are produced using the ModernBERT base model (further pretrained with the Nomic Embed framework) as an imputation function (Warner et al., 2024; Nussbaum et al., 2024).

Figure 5 shows that regardless of the annotation score π chosen, and its induced $\hat{\mu}$, point estimates appear unbiased, as we would expect given Assumption 3 (Wager, 2024). The more

data annotated, the smaller the confidence intervals, reflecting accuracy improvements in the classifier when more labeled training data are available. With a highly performant deep neural classifier like ModernBERT, a researcher need annotate only a small fraction of their overall dataset to produce estimates with precision comparable to labeling considerably more data (in this case, 403 versus 4157 labels for $\pi = 0.1$).

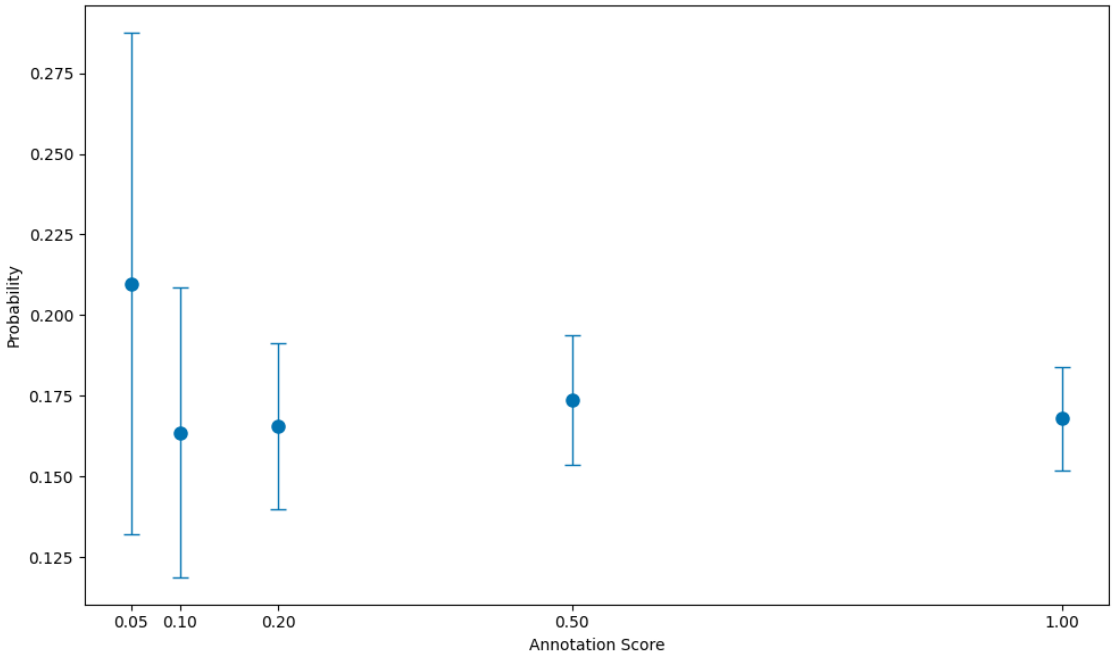


Figure 5: Estimating P(discussion of politics) with annotation score $\pi \in \left\{\frac{1}{20}, \frac{1}{10}, \frac{1}{5}, \frac{1}{2}, 1\right\}$.

Figure 6 explicitly plots imputation function accuracy against downstream point estimates and confidence intervals for a range of popular text classification models and methods. Specifically, we inspect performance and estimates based on three zero-shot text classification methods (the ChatGPT-informed keyword-classification strategy employed in Dell et al. (2023); a zero-shot encoder transformer per Yin et al. (2019); and a popular, open-source "GenAI" model, the smallest Llama 3.1 instruction-tuned decoder transformer (Grattafiori et al., 2024)) and two neural encoder methods (the original BERT encoder transformer (Devlin et al., 2019) and its recent successor ModernBERT (Warner et al., 2024) as pretrained with the Nomic Embed framework (Nussbaum et al., 2024)) fine-tuned with a dataset where $\pi = 0.2$.

Figure 4 shows that the zero-shot methods, including a popular open-source Llama model, offer noticeably worse performance for this simple topic classification exercise than supervised, fine-tuned methods. This worse performance in turn translates into significantly wider confidence intervals. A frequent question in the emerging work using deep learning to impute missing data in economics is how accurate is accurate enough? In the presence of non-classical measurement error—which is inherent when a deep neural network is the imputation function—the answer will depend on the precision that the researcher requires when using the imputed data to compute downstream estimators.

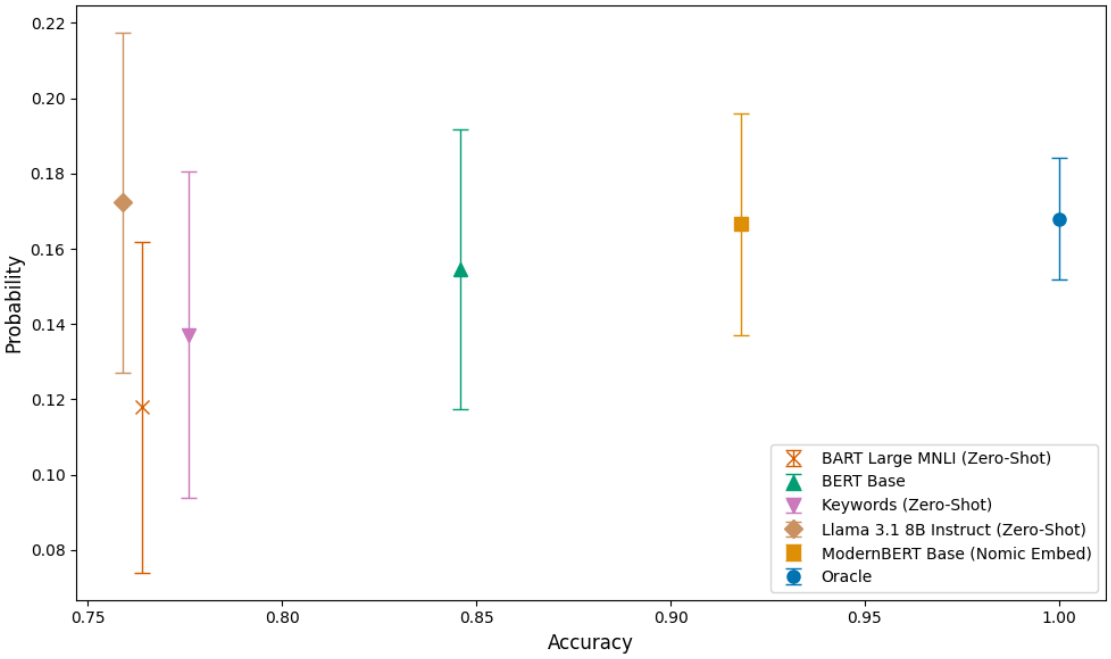


Figure 6: Estimating P(discussion of politics) across fine-tuned and zero-shot classifiers.

8 Conclusion

Unstructured data can generate valuable new insights into social science questions. However, incorporating structured information imputed from unstructured data into estimators, without accounting for the fact that the data are imputed with inherently non-classical measurement error, can lead to biased and inconsistent results, invalidating asymptotic uncertainty quantification.

By conceptualizing the statistical role of unstructured data as imputing missing structured data, the MARS framework is able to recover $\sqrt{n}$-consistent estimators along with valid uncertainty quantification. This further unlocks recipes for developing not just valid but also efficient and robust estimators and recovers estimators from related frameworks as (potentially inefficient) special cases. MARS, through its lens of efficiency, also reveals that "optimal" imputation functions that asymptotically enable efficient estimates can take non-obvious forms, as they require capturing variation not just in missing structured and unstructured data, but also variation in context-specific covariates.

Deriving efficient influence functions can be challenging for certain functionals. However, incorporating recent advances in automatic debiasing (Chernozhukov et al., 2022b; van der Laan et al., 2025) or automatic (functional) differentiation (Luedtke, 2024) into MARS could significantly reduce the burden on researchers. These methods and tools offer new opportunities to make MARS more user-friendly and scalable.

Casting efficient and robust inference with unstructured data as efficient and robust inference with missing structured data requires that a researcher define and measure the low-dimensional summaries that they wish to extract from unstructured data. Even though the imputation function (often a deep neural network) may be a black box, its output, and that output's role in the estimator, must be interpretable under the MARS framework. If a researcher leverages

unstructured data for unsupervised data exploration - where the objective is an open-ended exploration of patterns in the data and there is no single ground truth - MARS is not the appropriate framework.

We show that MARS can be readily applied. Its estimates can differ substantially from estimates produced without MARS, due to systematic measurement error. Our examples show that better prediction matters for statistical inference, leading to tighter confidence intervals. Importantly, all of the imputation functions trained and implemented in this paper were both highly performant and computationally lightweight, requiring minimal compute. If readers take one point away from this study, we hope that it will be that systematic measurement error - which is inherent to the imputation of structured data with neural networks - must be taken seriously. MARS offers an easy-to-implement framework that, under relatively weak assumptions, enables unbiased and efficient inference using imputed structured data.

References

Angelopoulos, A. N., S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic (2023): "Prediction-powered inference," *Science*, 382, 669–674, publisher: American Association for the Advancement of Science.

Angelopoulos, A. N., J. C. Duchi, and T. Zrnic (2024): "PPI++: Efficient Prediction-Powered Inference," ArXiv:2311.01453.

Athey, S., R. Chetty, G. Imbens, and H. Kang (2019): "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely," Tech. Rep. w26463, National Bureau of Economic Research, Cambridge, MA.

Bach, P., O. Schacht, V. Chernozhukov, S. Klaassen, and M. Spindler (2024): "Hyperparameter Tuning for Causal Inference with Double Machine Learning: A Simulation Study," ArXiv:2402.04674 [econ].

Baker, S. R., N. Bloom, and S. J. Davis (2016): "Measuring Economic Policy Uncertainty*," *The Quarterly Journal of Economics*, 131, 1593–1636.

Bang, H. and J. M. Robins (2005): "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–973.

Battaglia, L., T. Christensen, S. Hansen, and S. Sacher (2024): "Inference for Regression with Variables Generated by AI or Machine Learning," ArXiv:2402.15585 [econ].

Beltagy, I., M. E. Peters, and A. Cohan (2020): "Longformer: The Long-Document Transformer," ArXiv:2004.05150 [cs].

Ben-David, S., J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan (2010): "A theory of learning from different domains," *Machine learning*, 79, 151–175.

Bibaut, A. F. and M. J. van der Laan (2017): "Data-adaptive smoothing for optimal-rate estimation of possibly non-regular parameters," ArXiv:1706.07408 [math].

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1998): *Efficient and adaptive estimation for semiparametric models*, New York: Springer.

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022): "When is TSLS Actually LATE?" .

Cai, T. T. and H. Pu (2024): "Transfer Learning for Nonparametric Regression: Non-asymptotic Minimax Analysis and Adaptive Procedure," ArXiv:2401.12272.

Caldara, D. and M. Iacoviello (2022): "Measuring Geopolitical Risk," *American Economic Review*, 112, 1194–1225.

Callaway, B. and P. H. C. Sant'Anna (2021): "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 225, 200–230.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011): "Robust Inference With Multiway Clustering," *Journal of Business & Economic Statistics*, 29, 238–249.

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2024): "A Practical Introduction to Regression Discontinuity Designs: Extensions," *Elements in Quantitative and Computational Methods for the Social Sciences*, iSBN: 9781009441896 9781009462327 9781009441902 Publisher: Cambridge University Press.

Cattaneo, M. D. and R. Titiunik (2022): "Regression Discontinuity Designs," *Annual Review of Economics*, 14, 821–851, publisher: Annual Reviews.

Chapelle, O., B. Schölkopf, and A. Zien, eds. (2010): *Semi-supervised learning*, Adaptive computation and machine learning, Cambridge, Mass: MIT Press, 1st mit press pbk. ed ed.

Chen, Q., V. Syrgkanis, and M. Austern (2024): "Debiased machine learning without sample-splitting for stable estimators," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., NIPS '22, 3096–3109.

Chen, X., H. Hong, and D. Nekipelov (2011): "Nonlinear Models of Measurement Errors," *Journal of Economic Literature*, 49, 901–937.

Chen, X., H. Hong, and E. Tamer (2005): "Measurement Error Models with Auxiliary Data," *The Review of Economic Studies*, 72, 343–366.

Chen, X. and A. Santos (2018): "Overidentification in Regular Models," *Econometrica*, 86, 1771–1817, publisher: [Wiley, The Econometric Society].

Cheng, H.-T., L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, R. Anil, Z. Haque, L. Hong, V. Jain, X. Liu, and H. Shah (2016): "Wide & Deep Learning for Recommender Systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, New York, NY, USA: Association for Computing Machinery, DLRS 2016, 7–10.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68.

Chernozhukov, V., W. Newey, V. M. Quintas-Martinez, and V. Syrgkanis (2022a): "RieszNet and ForestRiesz: Automatic Debiased Machine Learning with Neural Nets and Random Forests," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, 3901–3914, iSSN: 2640-3498.

Chernozhukov, V., W. K. Newey, and R. Singh (2022b): "Automatic Debiased Machine Learning of Causal and Structural Effects," *Econometrica*, 90, 967–1027, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA18515.

Cytrynbaum, M. (2021): "Optimal Stratification of Survey Experiments," Version Number: 2.

Deaton, A. (1985): "Panel data from time series of cross-sections," *Journal of Econometrics*, 30, 109–126.

Dell, M., J. Carlson, T. Bryan, E. Silcock, A. Arora, Z. Shen, L. D'Amico-Wong, Q. Le, P. Querubin, and L. Heldring (2023): "American Stories: A Large-Scale Structured Text Dataset of Historical U.S. Newspapers," *Advances in Neural Information Processing Systems*, 36, 80744–80772.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019): "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, ed. by J. Burstein, C. Doran, and T. Solorio, Minneapolis, Minnesota: Association for Computational Linguistics, 4171–4186.

Ding, P. and F. Li (2018): "Causal Inference: A Missing Data Perspective," *Statistical Science*, 33, 214–237, publisher: Institute of Mathematical Statistics.

Drews, S. and M. Kohler (2024): "On the universal consistency of an over-parametrized deep neural network estimate learned by gradient descent," *Annals of the Institute of Statistical Mathematics*, 76, 361–391.

D'Amour, A., P. Ding, A. Feller, L. Lei, and J. Sekhon (2021): "Overlap in observational studies with high-dimensional covariates," *Journal of Econometrics*, 221, 644–654.

Egami, N., M. Hinck, B. Stewart, and H. Wei (2023): "Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models," *Advances in Neural Information Processing Systems*, 36, 68589–68601.

Egami, N., M. Hinck, B. M. Stewart, and H. Wei (2024): "Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses," .

Fuller, W. A. (1987): *Measurement error models*, Wiley series in probability and mathematical statistics. Probability and mathematical statistics, New York: Wiley.

Gibson, J., S. Olivia, and G. Boe-Gibson (2020): "Night lights in economics: Sources and uses," *Journal of Economic Surveys*, 34, 955–980.

Goodfellow, I., Y. Bengio, and A. Courville (2016): *Deep learning*, Adaptive computation and machine learning, Cambridge, Massachusetts: The MIT Press.

Grattafiori, A., A. Dubey, A. Jauhri, and the Llama team (2024): "The Llama 3 Herd of Models," ArXiv:2407.21783 [cs].

Hines, O., O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt (2022): "Demystifying Statistical Learning Based on Efficient Influence Functions," *The American Statistician*, 76, 292–304.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00442.

Ichimura, H. and W. K. Newey (2022): "The influence function of semiparametric estimators," *Quantitative Economics*, 13, 29–61, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE826.

Imbens, G. W. and D. B. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge: Cambridge University Press.

Ji, W., L. Lei, and T. Zrnic (2025): "Predictions as Surrogates: Revisiting Surrogate Outcomes in the Age of AI," Version Number: 1.

Kallenberg, O. (2021): *Foundations of Modern Probability*, vol. 99 of *Probability Theory and Stochastic Modelling*, Cham: Springer International Publishing.

Kallus, N. and X. Mao (2024): "On the role of surrogates in the efficient estimation of treatment effects with limited outcome data," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkae099.

Kennedy, E. H. (2016): "Semiparametric Theory and Empirical Processes in Causal Inference," in *Statistical Causal Inferences and Their Applications in Public Health Research*, ed. by H. He, P. Wu, and D.-G. D. Chen, Cham: Springer International Publishing, 141–167.

——— (2018): "Semiparametric Theory," in *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, 1–7, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat08083.

——— (2023): "Semiparametric doubly robust targeted double machine learning: a review," ArXiv:2203.06469.

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small (2017): "Nonparametric methods for doubly robust estimation of continuous treatment effects," *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 79, 1229–1245.

Klaassen, S., J. Teichert-Kluge, P. Bach, V. Chernozhukov, M. Spindler, and S. Vijaykumar (2024): "DoubleMLDeep: Estimation of Causal Effects with Multimodal Data," ArXiv:2402.01785.

Kluger, D. M., K. Lu, T. Zrnic, S. Wang, and S. Bates (2025): "Prediction-Powered Inference with Imputed Covariates and Nonuniform Sampling," Version Number: 1.

List, J. A., I. Muir, and G. Sun (2024): "Using machine learning for efficient flexible regression adjustment in economic experiments," *Econometric Reviews*, 0, 1–39, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07474938.2024.2373446.

Little, R. and D. Rubin (2019): *Statistical Analysis with Missing Data, Third Edition*, Wiley Series in Probability and Statistics, Wiley, 1 ed.

Liu, L., R. Mukherjee, W. K. Newey, and J. M. Robins (2023): "Semiparametric Efficient Empirical Higher Order Influence Function Estimators," ArXiv:1705.07577 [math].

Ludwig, J., S. Mullainathan, and A. Rambachan (2024): "Large Language Models: An Applied Econometric Framework," ArXiv:2412.07031 [econ].

Luedtke, A. (2024): "Simplifying debiased inference via automatic differentiation and probabilistic programming," ArXiv:2405.08675 [stat].

Ma, Y., P. H. C. Sant'Anna, Y. Sasaki, and T. Ura (2023): "Doubly Robust Estimators with Weak Overlap," ArXiv:2304.08974 [econ].

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382, publisher: [Wiley, Econometric Society].

Neyman, J. (1923): "On the application of probability theory to agricultural experiments. Essay on principles," *Ann. Agricultural Sciences*, 1–51.

Noack, C., T. Olma, and C. Rothe (2024): "Flexible Covariate Adjustments in Regression Discontinuity Designs," ArXiv:2107.07942 [econ].

Nussbaum, Z., J. X. Morris, B. Duderstadt, and A. Mulyar (2024): "Nomic Embed: Training a Reproducible Long Context Text Embedder," ArXiv:2402.01613 [cs].

Pfanzagl, J. (1982): *Contributions to a General Asymptotic Statistical Theory*, vol. 13 of *Lecture Notes in Statistics*, New York, NY: Springer New York.

Rambachan, A., R. Singh, and D. Viviano (2024a): "Program Evaluation with Remotely Sensed Outcomes," *arXiv preprint arXiv:2411.10959*.

——— (2024b): "Program Evaluation with Remotely Sensed Outcomes," ArXiv:2411.10959 [econ].

Reimers, N. and I. Gurevych (2019): "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, ed. by K. Inui, J. Jiang, V. Ng, and X. Wan, Hong Kong, China: Association for Computational Linguistics, 3982–3992.

Robins, J., L. Li, E. Tchetgen, and A. van der Vaart (2008): "Higher order influence functions and minimax estimation of nonlinear functionals," in *Probability and Statistics: Essays in Honor of David A. Freedman*, Institute of Mathematical Statistics, vol. 2, 335–422.

Robins, J. M., L. Li, R. Mukherjee, E. T. Tchetgen, and A. v. d. Vaart (2017): "Minimax estimation of a functional on a structured high-dimensional model," *The Annals of Statistics*, 45, 1951–1987, publisher: Institute of Mathematical Statistics.

Robins, J. M. and A. Rotnitzky (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90, 122–129, publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1995.10476494.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866, publisher: [American Statistical Association, Taylor & Francis, Ltd.].

——— (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121, publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1995.10476493.

Rotnitzky, A., J. M. Robins, and D. O. Scharfstein (1998): "Semiparametric Regression for Repeated Outcomes with Nonignorable Nonresponse," *Journal of the American Statistical Association*, 93, 1321–1339, publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Rubin, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, 66, 688.

——— (1976): "Inference and Missing Data," *Biometrika*, 63, 581–592, publisher: [Oxford University Press, Biometrika Trust].

——— (1978): "Bayesian inference for causal effects: The role of randomization," *The Annals of statistics*, 34–58.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999): "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, 94, 1096–1120, publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Schennach, S. (2022): "Measurement Systems," *Journal of Economic Literature*, 60, 1223–1263.

Schennach, S. M. (2016): "Recent Advances in the Measurement Error Literature," *Annual Review of Economics*, 8, 341–377.

Schuler, A. and M. van der Laan (2024): *Introduction to Modern Causal Inference*.

Semenova, V. and V. Chernozhukov (2021): "Debiased machine learning of conditional average treatment effects and other causal functions," *The Econometrics Journal*, 24, 264–289.

Sigman, K. (2010): "Professor Karl Sigman's Lecture Notes on Monte Carlo Simulation," .

Stone, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053, publisher: Institute of Mathematical Statistics.

Tsiatis, A. (2006): *Semiparametric Theory and Missing Data*, Springer Series in Statistics, New York, NY: Springer New York.

van der Laan, L., A. Bibaut, N. Kallus, and A. Luedtke (2025): "Automatic Debiased Machine Learning for Smooth Functionals of Nonparametric M-Estimands," ArXiv:2501.11868 [stat].

van der Laan, M. J. and D. Rubin (2006): "Targeted Maximum Likelihood Learning," *The International Journal of Biostatistics*, 2, publisher: De Gruyter.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge: Cambridge University Press.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin (2017): "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 30.

Wager, S. (2024): *Causal Inference: A Statistical Learning Approach*.

Wang, S., S. Di Tommaso, J. M. Deines, and D. B. Lobell (2020): "Mapping twenty years of corn and soybean across the US Midwest using the Landsat archive," *Scientific Data*, 7, 307.

Warner, B., A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, and I. Poli (2024): "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference," ArXiv:2412.13663 [cs].

Yang, D., Y.-H. H. Tsai, and M. Yamada (2024): "On Verbalized Confidence Scores for LLMs," Version Number: 1.

Yin, W., J. Hay, and D. Roth (2019): "Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, ed. by K. Inui, J. Jiang, V. Ng, and X. Wan, Hong Kong, China: Association for Computational Linguistics, 3914–3923.

Zhang, Y. and J. Bradic (2022): "High-dimensional semi-supervised learning: in search of optimal inference of the mean," *Biometrika*, 109, 387–403.

Zhang, Y., A. Chakrabortty, and J. Bradic (2023): "Double Robust Semi-Supervised Inference for the Mean: Selection Bias under MAR Labeling with Decaying Overlap," ArXiv:2104.06667.

Zrnic, T. and E. Candes (2024): "Active Statistical Inference," in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, 62993–63010, iSSN: 2640-3498.

Zrnic, T. and E. J. Candès (2024): "Cross-prediction-powered inference," *Proceedings of the National Academy of Sciences*, 121, e2322083121, publisher: Proceedings of the National Academy of Sciences.

9 Appendix

9.1 Proofs

9.1.1 Proposition 1

We define a MARS mean functional as any functional that can be written as

$$\theta(P) = E_P[\tilde{M}^*]$$

where $\tilde{M}^* = g(M^*, V)$ and where g is a known deterministic function and V is a known random variable, $[A \perp\!\!\!\perp V] \mid X$. When g is homogeneous of degree 1 in its first argument, under Assumption 1 we have that

$$A\tilde{M}^* = Ag(M^*, V) = g(AM^*, V) = g(M, V) := \tilde{M}.$$

Under Assumption 2, we have that

$$[M^* \perp\!\!\!\perp A] \mid (X, U)$$

and so

$$[g(M^*, V) \perp\!\!\!\perp A] \mid (X, U, V).$$

Labeling $\tilde{X} = (X, U, V)$, we can also write this as

$$[\tilde{M}^* \perp\!\!\!\perp A] \mid \tilde{X}.$$

Specifically, then, being a MARS mean functional means that point identification can always be achieved as

$$\begin{aligned}
\theta &= E[\tilde{M}^*] \\
&= E[E[\tilde{M}^* \mid \tilde{X}]] \\
&= E[E[\tilde{M}^* \mid A = 1, \tilde{X}]] && \text{(Assumption 2)} \\
&= E[E[\tilde{M} \mid A = 1, \tilde{X}]] && \text{(Assumption 1)} \\
&= E[\tilde{\mu}(\tilde{X})]
\end{aligned}$$

where $\tilde{\mu}(\tilde{X}) := E[\tilde{M} \mid A = 1, \tilde{X}]$. Explicitly, we can write

$$\theta = \int_{\tilde{\mathcal{X}}} \left(\int_{\tilde{\mathcal{M}}} \tilde{m}\, p(\tilde{m} \mid 1, \tilde{x})\, d\tilde{m} \right) p(\tilde{x})\, d\tilde{x}.$$

Note that all of the functionals involving missing structured data considered in section 4 are MARS mean functionals. This encompasses conditional MARS mean functionals, such as for the DiD and RDD applications, which can be re-written as

$$\theta = E[M^* \mid V = v] = E\left[\frac{\mathbb{I}\{V = v\}}{P(V = v)} M^* \right] = E[\tilde{M}^*],$$

with $\tilde{M}^* = \frac{\mathbb{I}\{V=v\}}{P(V=v)} M^*$, for some discrete conditioning variable V.

9.1.2 Proposition 2

Proof. The first part of the proof of this proposition is identical to the proof of Proposition 4, which delivers that the efficient influence function for functional $E_P\left[\tilde{\mu}(\tilde{X}_i)\right]$, under the semiparametric statistical model $\mathcal{P}_\pi \ni P$ induced by Assumption 3, is

$$\varphi(W_i) = \tilde{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(\tilde{X}_i)}(\tilde{M}_i - \tilde{\mu}(\tilde{X}_i)) - \theta.$$

Then, under Assumption 2, we notice that

$$\pi(\tilde{X}) := P(A = 1 \mid \tilde{X}) = P(A = 1 \mid X, U, V) = P(A = 1 \mid X, V) := \pi(X, V)$$

and so

$$\varphi(W_i) = \tilde{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i, V_i)}(\tilde{M}_i - \tilde{\mu}(\tilde{X}_i)) - \theta.$$

□

9.1.3 Proposition 3

We know from classic semiparametric theory that the semiparametric variance lower bound for $E_P\left[\tilde{\mu}(\tilde{X}_i)\right]$ will be given by $\mathrm{Var}(\varphi(W_i))$ in our setting, where, per Proposition 2,

$$\varphi = \tilde{\mu}(\tilde{X}) + \frac{A}{\pi(X, V)}(\tilde{M} - \tilde{\mu}(\tilde{X})) - \theta.$$

Classic results in semiparametric inference for missing data tell us that, with some careful algebra,

$$\begin{aligned}
\mathrm{Var}(\varphi) &= \mathrm{Var}\left(\tilde{\mu}(\tilde{X}) + \frac{A}{\pi(X, V)}(\tilde{M} - \tilde{\mu}(\tilde{X})) - \theta\right) \\
&= \mathrm{Var}\left(\tilde{\mu}(\tilde{X}) + \frac{A}{\pi(X, V)}(\tilde{M} - \tilde{\mu}(\tilde{X}))\right) \\
&= \mathrm{Var}\left(\tilde{\mu}(\tilde{X})\right) + \mathrm{Var}\left(\frac{A}{\pi(X, V)}(\tilde{M} - \tilde{\mu}(\tilde{X}))\right) \\
&= \mathrm{Var}(E[\tilde{M}^* \mid \tilde{X}]) + E\left[\frac{\mathrm{Var}(\tilde{M}^* \mid \tilde{X})}{\pi(X, V)}\right].
\end{aligned}$$

Then, by the law of total variance, we have that

$$\begin{aligned}
\mathrm{Var}(\varphi) &= \mathrm{Var}(E[\tilde{M}^* \mid \tilde{X}]) + E\left[\frac{\mathrm{Var}(\tilde{M}^* \mid \tilde{X})}{\pi(X,V)}\right] \\
&= \mathrm{Var}(M^*) - E[\mathrm{Var}(\tilde{M}^* \mid \tilde{X})] + E\left[\frac{\mathrm{Var}(\tilde{M}^* \mid \tilde{X})}{\pi(X,V)}\right] \\
&= \mathrm{Var}(M^*) + E\left[\frac{\mathrm{Var}(\tilde{M}^* \mid \tilde{X})}{\pi(X,V)} - \mathrm{Var}(\tilde{M}^* \mid \tilde{X})\right] \\
&= \mathrm{Var}(M^*) + E\left[\frac{\mathrm{Var}(\tilde{M}^* \mid \tilde{X})}{\pi(X,V)} - \mathrm{Var}(\tilde{M}^* \mid \tilde{X})\right] \\
&= \mathrm{Var}(\tilde{M}^*) + E[(\pi(X,V)^{-1} - 1)\mathrm{Var}(\tilde{M}^* \mid \tilde{X})].
\end{aligned}$$

Then, under a constant annotation score π,

$$\begin{aligned}
\mathrm{Var}(\varphi) &= \mathrm{Var}(\tilde{M}^*) + E[(\pi(X,V)^{-1} - 1)\mathrm{Var}(\tilde{M}^* \mid \tilde{X})] \\
&= \mathrm{Var}(\tilde{M}^*) + E[(\pi^{-1} - 1)\mathrm{Var}(\tilde{M}^* \mid \tilde{X})] \\
&= \mathrm{Var}(\tilde{M}^*) + (\pi^{-1} - 1)E[\mathrm{Var}(\tilde{M}^* \mid \tilde{X})] \\
&= \mathrm{Var}(\tilde{M}^*) + (\pi^{-1} - 1)\,\mathrm{Var}(\tilde{M}^*)(1 - R^2(\tilde{X})) \\
&= \mathrm{Var}(\tilde{M}^*)\left(1 + \left(\frac{1}{\pi} - 1\right)\left(1 - R^2(\tilde{X})\right)\right),
\end{aligned}$$

which uses the fact that $E[\mathrm{Var}(\tilde{M}^* \mid \tilde{X})] := \mathrm{Var}(\tilde{M}^*)(1 - R^2(\tilde{X}))$, where $R^2(\tilde{X})$ the nonparametric R^2.

9.1.4 Lemma 1

Proof. Let us first assume that $\mathcal{P} \ni P$ is a fully nonparametric statistical model, which is to say Assumption 3 does not hold. Under $\mathcal{P}$, without any restrictions, P generically factors as:

$$P(\tilde{M}, A, \tilde{X}) = P(\tilde{M} \mid A, \tilde{X})P(A \mid \tilde{X})P(\tilde{X}).$$

The tangent space $\mathcal{T}$ (i.e., the set of all valid scores h associated with P and $\mathcal{P}$) for this model can be constructed as the the orthogonal sum of the tangent spaces for each factorized component of the model:

$$\begin{aligned}
\mathcal{T} &= \mathcal{T}_{\tilde{M}|A,\tilde{X}} \oplus \mathcal{T}_{A|\tilde{X}} \oplus \mathcal{T}_{\tilde{X}} \\
\mathcal{T}_{\tilde{X}} &= \{h_{\tilde{X}}(\tilde{x}) \in L^2(P) : E_P[h_{\tilde{X}}(\tilde{X})] = 0\} \\
\mathcal{T}_{A|\tilde{X}} &= \{h_{A|\tilde{X}}(a, \tilde{x}) \in L^2(P) : E_P[h_{A|\tilde{X}}(A, \tilde{X}) \mid \tilde{X}] = 0\} \\
\mathcal{T}_{\tilde{M}|A,\tilde{X}} &= \{h_{\tilde{M}|A,\tilde{X}}(\tilde{m}, a, \tilde{x}) \in L^2(P) : E_P[h_{\tilde{M}|A,\tilde{X}}(\tilde{M}, A, \tilde{X}) \mid A, \tilde{X}] = 0\}.
\end{aligned}$$

A score h under this nonparametric model can therefore be factorized as

$$h(\tilde{M}, A, \tilde{X}) = h_{\tilde{M}|A,\tilde{X}}(\tilde{M}, A, \tilde{X}) + h_{A|\tilde{X}}(A, \tilde{X}) + h_{\tilde{X}}(\tilde{X}),$$

where all summands are pairwise orthogonal.

Denoting a pathwise derivative with respect to h as ∇_h, we then have that, by linearity of pathwise derivatives,

$$\nabla_h \theta = \nabla_{h_{\tilde{M}|A,\tilde{X}}} \theta + \nabla_{h_{A|\tilde{X}}} \theta + \nabla_{h_{\tilde{X}}} \theta.$$

Crucially, notice then that $\nabla_{h_{A|\tilde{X}}} \theta = 0$ for any score h at P:

$$\begin{aligned}
\nabla_{h_{A|\tilde{X}}} \theta &= \frac{d}{d\epsilon} \theta(P_{\epsilon, h_{A|\tilde{X}}}) \\
&= \frac{d}{d\epsilon} \int_{\tilde{\mathcal{X}}} \left(\int_{\tilde{\mathcal{M}}} \tilde{m}\, p(\tilde{m} \mid 1, \tilde{x})\, d\tilde{m} \right) p(\tilde{x})\, d\tilde{x} \\
&= 0.
\end{aligned}$$

As such, we have that

$$\nabla_h \theta = \nabla_{h_{\tilde{M}|A,\tilde{X}}} \theta + \nabla_{h_{\tilde{X}}} \theta$$

for any score h.

Alternatively, if we had made Assumption 3, then our statistical model would be semiparametric, which we can denote $\mathcal{P}_\pi$, where the distribution factorizes as

$$P(\tilde{M}, A, \tilde{X}) = P(\tilde{M} \mid A, \tilde{X})(A\pi(\tilde{X}) + (1 - A)(1 - \pi(\tilde{X})))P(\tilde{X})$$

for known annotation score function π. For $\mathcal{P}_\pi$ and $P \in \mathcal{P}_\pi$, the corresponding tangent space $\mathcal{T}_\pi$ therefore factorizes as

$$\mathcal{T}_\pi = \mathcal{T}_{\tilde{M}|A,\tilde{X}} \oplus \mathcal{T}_{\tilde{X}}.$$

Any score h_π under this semiparametric model can therefore be factorized as

$$h_\pi(\tilde{M}, A, \tilde{X}) = h_{\tilde{M}|A,\tilde{X}} + h_{\tilde{X}},$$

where the summands are again orthogonal. Again by the linearity of the pathwise derivative operator, we have that

$$\nabla_{h_\pi} \theta = \nabla_{h_{\tilde{M}|A,\tilde{X}}} \theta + \nabla_{h_{\tilde{X}}} \theta.$$

As such, we have learned that

$$\boxed{\nabla_h \theta = \nabla_{h_{\tilde{M}|A,\tilde{X}}} \theta + \nabla_{h_{\tilde{X}}} \theta = \nabla_{h_\pi} \theta.}$$

By the definition of pathwise differentiability, we know that

$$\nabla_{h_\pi} \theta = E[\varphi_\pi h_\pi]$$

holds, where φ_π is the efficient influence function for $\theta(P)$ under $P \in \mathcal{P}_\pi$, by the Riesz representation theorem. However, if φ_π is the EIF under $\mathcal{P}_\pi$, then it lies in the tangent space $\mathcal{T}_\pi$,

which is orthogonal to $\mathcal{T}_{A|\tilde{X}}$ and all of the scores $h_{A|\tilde{X}}$ in it. This means we can write

$$\begin{aligned}\nabla_{h_\pi}\theta &= E[\varphi_\pi h_\pi] + 0 \\ &= E[\varphi_\pi h_\pi] + E[\varphi_\pi h_{A|\tilde{X}}] \\ &= E[\varphi_\pi (h_\pi + h_{A|\tilde{X}})] \\ &= E[\varphi_\pi h].\end{aligned}$$

Putting everything together then, we have that

$$\nabla_h \theta = \nabla_{h_\pi}\theta = E[\varphi_\pi h].$$

Importantly then,

$$\boxed{\nabla_h \theta = E[\varphi_\pi h]}$$

which, again by the definition of pathwise differentiability, tells us that the EIF for the semiparametric model $\mathcal{P}_\pi$ induced by Assumption 3, which is φ_π, is the same as the nonparametric EIF, which is the random variable that would satisfy this equality.

As a result, we conclude that it suffices to find the EIF for a given MARS mean functional $\theta(P)$ under a nonparametric model $\mathcal{P}$ in order to find the EIF for $\theta(P)$ under the semiparametric model $\mathcal{P}_\pi$ induced by Assumption 3. □

9.1.5 Proposition 4

Proof. In order to derive the efficient influence function for functional $\theta(P) = E_P\left[\mu(\tilde{X})\right]$ under the semiparametric statistical model $\mathcal{P}_\pi \ni P$ induced by Assumption 3, we first first find the EIF for $E_P\left[\mu(\tilde{X})\right]$ under a nonparametric statistical model $\mathcal{P} \ni P$, and then apply Lemma 1.

To find the nonparametric EIF for the functional of interest, we will (1) find a candidate EIF by treating the all data as discrete and making use of the influence function operator (Kennedy, 2023), and then (2) check this is indeed a valid influence function in the general, non-discrete case using the definition of pathwise differentiability.

Treating all data as discrete, application of the influence function operator to our functional

of interest yields:

$$\begin{aligned}
\mathbb{IF}(E_P\left[\mu(\tilde{X})\right]) &= \mathbb{IF}(\sum_{\tilde{x}\in\tilde{\mathcal{X}}}\mu(\tilde{x})P(\tilde{X}=\tilde{x})) \\
&= \sum_{\tilde{x}\in\tilde{\mathcal{X}}}\mathbb{IF}(\mu(\tilde{x})P(\tilde{X}=\tilde{x})) \\
&= \sum_{\tilde{\mathcal{X}}}\left[\mathbb{IF}(\mu(\tilde{x}))P(\tilde{X}=\tilde{x}) + \mu(\tilde{x})\mathbb{IF}(P(\tilde{X}=\tilde{x}))\right] \\
&= \sum_{\tilde{x}\in\tilde{\mathcal{X}}}\left[\frac{\mathbb{I}\{\tilde{X}=\tilde{x}, A=1\}}{P(\tilde{X}=\tilde{x}, A=1)}(M-\mu(\tilde{x}))P(\tilde{X}=\tilde{x}) + \mu(\tilde{x})(\mathbb{I}\{\tilde{X}=\tilde{x}\} - P(\tilde{X}=\tilde{x}))\right] \\
&= \sum_{\tilde{x}\in\tilde{\mathcal{X}}}\left[\frac{A\mathbb{I}\{\tilde{X}=\tilde{x}\}}{P(\tilde{X}=\tilde{x}, A=1)}(M-\mu(\tilde{x}))P(\tilde{X}=\tilde{x}) + \mu(\tilde{x})\mathbb{I}\{\tilde{X}=\tilde{x}\} - \mu(\tilde{x})P(\tilde{X}=\tilde{x})\right] \\
&= \frac{A}{P(\tilde{X}, A=1)}(M-\mu(\tilde{X}))P(\tilde{X}) + \mu(\tilde{X}) - E\left[\mu(\tilde{X})\right] \\
&= \frac{A}{P(A=1\mid\tilde{X})}(M-\mu(\tilde{X})) + \mu(\tilde{X}) - E\left[\mu(\tilde{X})\right] \\
&= \frac{A}{\pi(\tilde{X})}(M-\mu(\tilde{X})) + \mu(\tilde{X}) - \theta.
\end{aligned}$$

Rearranging, our candidate EIF φ (which would be the nonparametric EIF if all data were discrete) is

$$\varphi := \mu(\tilde{X}) + \frac{A}{\pi(\tilde{X})}(M-\mu(\tilde{X})) - \theta.$$

We now apply the definition of pathwise differentiability to check that this is indeed the (efficient) influence function for θ. In particular, the definition of pathwise differentiability tells us that

$$\nabla_h\theta = E[\varphi h].$$

For the left-hand side, we have that

$$\nabla_h\theta = \frac{d}{d\epsilon}\int_{\tilde{\mathcal{X}}}\left(\int_{\mathcal{M}} m(1+\epsilon h)p(m\mid 1,\tilde{x})\,dm\right)(1+\epsilon h)p(\tilde{x})\,d\tilde{x}$$

and for the right-hand side, we have that

$$\begin{aligned}
E[\varphi h] &= E\left[\left(\mu(\tilde{X}) + \frac{A}{\pi(\tilde{X})}(M - \mu(\tilde{X})) - \theta\right) h\right] \\
&= E\left[\left(\mu(\tilde{X}) + \frac{A}{\pi(\tilde{X})}(M - \mu(\tilde{X}))\right) h\right] - \theta E[h] \\
&= E\left[\left(\mu(\tilde{X}) + \frac{A}{\pi(\tilde{X})}(M - \mu(\tilde{X}))\right) h\right] \\
&= E\left[\mu(\tilde{X})h\right] + E\left[\frac{A}{\pi(\tilde{X})}(M - \mu(\tilde{X}))h\right] \\
&= E\left[\mu(\tilde{X})E[h \mid \tilde{X}]\right] + E\left[E\left[(M - \mu(\tilde{X}))h\frac{A}{\pi(\tilde{X})} \mid \tilde{X}\right]\right] \\
&= E\left[\mu(\tilde{X})E[h \mid \tilde{X}]\right] + E\left[E\left[(M - \mu(\tilde{X}))h\frac{1}{\pi(\tilde{X})} \mid \tilde{X}, A = 1\right] P(A = 1 \mid \tilde{X})\right] \\
&= E\left[\mu(\tilde{X})E[h \mid \tilde{X}]\right] + E\left[E\left[(M - \mu(\tilde{X}))h \mid \tilde{X}, A = 1\right]\right].
\end{aligned}$$

As discussed in Lemma 1, we know that in a fully nonparametric model $\mathcal{P}$ we can factorize the score as

$$h(M, A, \tilde{X}) = h_{M|A,\tilde{X}}(M, A, \tilde{X}) + h_{A|\tilde{X}}(A, \tilde{X}) + h_{\tilde{X}}(\tilde{X}).$$

As such, we have that

$$E[\varphi h] = E[\varphi h_{M|A,\tilde{X}}] + E[\varphi h_{A|\tilde{X}}] + E[\varphi h_{\tilde{X}}]$$

and by linearity of the pathwise derivative we also have that

$$\nabla_h \theta = \nabla_{h_{M|A,\tilde{X}}} \theta + \nabla_{h_{A|\tilde{X}}} \theta + \nabla_{h_{\tilde{X}}} \theta.$$

Checking equality summand by summand, we have:

- $h_{\tilde{X}}$ direction:

$$\begin{aligned}
E[\varphi h_{\tilde{X}}(\tilde{X})] &= E\left[\mu(\tilde{X})h_{\tilde{X}}(\tilde{X})\right] + E\left[E\left[(M - \mu(\tilde{X})) \mid \tilde{X}, A = 1\right] h_{\tilde{X}}(\tilde{X})\right] = E\left[\mu(\tilde{X})h_{\tilde{X}}(\tilde{X})\right] \\
\nabla_{h_{\tilde{X}}} \theta &= \int_{\tilde{\mathcal{X}}} \left(\int_{\mathcal{M}} mp(m \mid 1, \tilde{x})\, dm\right) h_{\tilde{X}}(\tilde{x}) p(\tilde{x})\, d\tilde{x} = E[\mu(\tilde{X})h_{\tilde{X}}(\tilde{X})]
\end{aligned}$$

  and so $\nabla_{h_{\tilde{X}}} \theta = E[\varphi h_{\tilde{X}}(\tilde{X})] = E[\mu(\tilde{X})h_{\tilde{X}}(\tilde{X})]$.

- $h_{A|\tilde{X}}$ direction:

$$\begin{aligned}
E[\varphi h_{A|\tilde{X}}] &= E\left[\mu(\tilde{X})E[h_{A|\tilde{X}} \mid \tilde{X}]\right] + E\left[E\left[(M - \mu(\tilde{X}))h_{A|\tilde{X}} \mid \tilde{X}, A = 1\right]\right] \\
&= E\left[E\left[(M - \mu(\tilde{X})) \mid \tilde{X}, A = 1\right] h_{A|\tilde{X}}\right] = 0 \\
\nabla_{h_{A|\tilde{X}}} \theta &= \frac{d}{d\epsilon} \int_{\tilde{\mathcal{X}}} \left(\int_{\mathcal{M}} mp(m \mid 1, \tilde{x})\, dm\right) p(\tilde{x})\, d\tilde{x} = 0
\end{aligned}$$

  and so $\nabla_{h_{A|\tilde{X}}} \theta = E[\varphi h_{A|\tilde{X}}] = 0$.

- $h_{M|A,\tilde{X}}$ direction:

$$\begin{aligned}
E[\varphi h_{M|A,\tilde{X}}] &= E\left[\mu(\tilde{X})E[h_{M|A,\tilde{X}} \mid \tilde{X}]\right] + E\left[E\left[(M-\mu(\tilde{X}))h_{M|A,\tilde{X}} \mid \tilde{X}, A=1\right]\right] \\
&= E\left[\mu(\tilde{X})E[h_{M|A,\tilde{X}} \mid \tilde{X}]\right] + E\left[E\left[Mh_{M|A,\tilde{X}} \mid \tilde{X}, A=1\right]\right] \\
&= E\left[\mu(\tilde{X})E[E[h_{M|A,\tilde{X}} \mid \tilde{X}, A] \mid \tilde{X}]\right] + E\left[E\left[Mh_{M|A,\tilde{X}} \mid \tilde{X}, A=1\right]\right] \\
&= E\left[E\left[Mh_{M|A,\tilde{X}} \mid \tilde{X}, A=1\right]\right] \\
\nabla_{h_{M|A,\tilde{X}}}\theta &= \frac{d}{d\epsilon}\int_{\tilde{\mathcal{X}}}\left(\int_{\mathcal{M}} m(1+\epsilon h_{M|A,\tilde{X}}(m,1,\tilde{x}))p(m \mid 1,\tilde{x})\,dm\right)p(\tilde{x})\,d\tilde{x} \\
&= \frac{d}{d\epsilon}\int_{\tilde{\mathcal{X}}}\left(\int_{\mathcal{M}} m\epsilon h_{M|A,\tilde{X}}(m,1,\tilde{x})p(m \mid 1,\tilde{x})\,dm\right)p(\tilde{x})\,d\tilde{x} \\
&= \int_{\tilde{\mathcal{X}}}\left(\int_{\mathcal{M}} m h_{M|A,\tilde{X}}(m,1,\tilde{x})p(m,1,\tilde{x})\,dm\right)p(\tilde{x})\,d\tilde{x} \\
&= E[E[Mh_{M|A,\tilde{X}} \mid A=1,\tilde{X}]]
\end{aligned}$$

and so $\nabla_{h_{M|A,\tilde{X}}}\theta = E[\varphi h_{M|A,\tilde{X}}] = E[E[Mh_{M|A,\tilde{X}} \mid A=1,\tilde{X}]]$.

Because all summands are equal, the definition of pathwise differentiability holds for the candidate φ, and so indeed

$$\varphi := \mu(\tilde{X}) + \frac{A}{\pi(\tilde{X})}(M-\mu(\tilde{X})) - \theta$$

is the influence function (or curve) for θ, and because $\mathcal{P}$ is nonparametric, it is also the EIF.

Letting $\tilde{M}^* := M^*$, by Lemma 1, we also know that φ is the EIF under the semiparametric model $\mathcal{P}_\pi$ induced by Assumption 3.

Finally, through Assumption 2, we have that

$$\pi(\tilde{X}) = \pi(X)$$

and so

$$\varphi := \mu(\tilde{X}) + \frac{A}{\pi(X)}(M-\mu(\tilde{X})) - \theta.$$

□

9.1.6 Proposition 5

Proof. We first want to show that the efficient influence function for the numerator functional $E_P\left[C_{i,j}^{\perp}\mu(\tilde{X}_i)\right]$ under the semiparametric statistical model $\mathcal{P}_\pi \ni P$ induced by Assumption 3 is

$$\varphi_{\text{num}}(W_i) = C_{i,j}^{\perp}\left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i))\right] - \theta_{j,\text{num}}.$$

Notice that, letting $\tilde{M}_i := C_{i,j}^{\perp}M_i$ and $\tilde{M}_i^* := C_{i,j}^{\perp}M_i^*$ and defining $\tilde{\mu}(\tilde{X}_i) := E[\tilde{M}_i \mid A_i = 1,\tilde{X}_i]$, then we can re-write our functional of interest simply as $E_P[\tilde{\mu}(\tilde{X}_i)]$, i.e., it is a MARS mean functional. Using identical arguments to Proposition 4, we know that the EIF for this

functional under a nonparametric statistical model is

$$\varphi_{\text{num}}(W_i) = \tilde{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(\tilde{X}_i)}(\tilde{M}_i - \tilde{\mu}(\tilde{X}_i)) - \theta_{j,\text{num}}.$$

Then, because the functional is a MARS mean functional, we apply Lemma 1 to conclude that it is also the EIF under $\mathcal{P}_\pi$. Applying Assumption 2 and recalling that $[C_i \perp\!\!\!\perp A_i] \mid X_i$ implies that $\pi(\tilde{X}_i) = \pi(X_i)$, which yields

$$\varphi_{\text{num}}(W_i) = \tilde{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(\tilde{M}_i - \tilde{\mu}(\tilde{X}_i)) - \theta_{j,\text{num}}.$$

We now want to prove that the efficient influence function for the denominator functional $E_P[{C_{i,j}^\perp}^2]$ is given by

$$\varphi_{\text{den}}(W_i) = {C_{i,j}^\perp}^2 - \theta_{j,\text{den}}.$$

To do so, we again (1) find a candidate EIF using basic calculus rules inherited by an influence function operator $\mathbb{IF}$, pretending the data are discrete and (2) check that this is indeed the EIF in a general, non-discrete case with the definition of pathwise differentiability. Because the denominator functional is just concerned with marginal distribution of the random variable ${C_{i,j}^\perp}^2$, we can effectively treat the relevant statistical model as nonparametric, as $\nabla_h \theta_{j,\text{den}} = \nabla_{h_{C_k^{\perp 2}}} \theta_{j,\text{den}} = \nabla_{h_\pi} \theta_{j,\text{den}}$.

To start, we have that

$$\varphi_{\text{den}} := \mathbb{IF}(E_P[{C_{i,j}^\perp}^2]) = {C_{i,j}^\perp}^2 - E_P[{C_{i,j}^\perp}^2].$$

The definition of pathwise differentiability thus requires that

$$\nabla_h \theta_{j,\text{den}} = \nabla_{h_{C_k^{\perp 2}}} \theta_{j,\text{den}} = E[\varphi_{\text{den}} h_{C_k^{\perp 2}}] = E[\varphi_{\text{den}} h],$$

where the last equality comes from the fact that the EIF under score $h_{C_k^{\perp 2}}$, per the marginal model, will be orthogonal to all other factored score functions.

Checking this, we have that

$$\begin{aligned}
\theta_{j,\text{den}} \nabla_{h_{C_k^{\perp 2}}} &= \frac{d}{d\epsilon} \int_{\mathcal{C}} c(1 + \epsilon h_{h_{C_k^{\perp 2}}}(c)) p(c)\, dc \\
&= \int_{\mathcal{C}} c h_{C_k^{\perp 2}}(c) p(c)\, dc \\
&= E[{C_k^\perp}^2 h_{C_k^{\perp 2}}]
\end{aligned}$$

and

$$\begin{aligned}
E[\varphi_{\text{den}} h_{C_k^{\perp 2}}] &= E[({C_{i,j}^\perp}^2 - E[{C_{i,j}^\perp}^2]) h_{C_k^{\perp 2}}] \\
&= E[{C_{i,j}^\perp}^2 h_{C_k^{\perp 2}}] - E[{C_{i,j}^\perp}^2] E[h_{C_k^{\perp 2}}] \\
&= E[{C_{i,j}^\perp}^2 h_{C_k^{\perp 2}}].
\end{aligned}$$

Therefore, indeed,

$$\varphi_{\text{den}}(W_i) = {C_{i,j}^{\perp}}^2 - \theta_{j,\text{den}}$$

is the EIF for the denominator functional. □

9.1.7 Proposition 6

Proof. We first want to show that the efficient influence function for the denominator functional $E_P[Z_i^{\perp}\mu(\tilde{X}_i)]$ under $\mathcal{P}_\pi \ni P$ is

$$\varphi_{\text{den}}(W_i) = Z_i^{\perp}\left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(\tilde{X}_i)}(M_i - \mu(\tilde{X}_i))\right] - \theta_{\text{den}}.$$

Much like the proof of Proposition 5, we may notice that, letting $\tilde{M}_i := Z_i^{\perp}M_i$ and $\tilde{M}_i^* := Z_i^{\perp}M_i^*$, and defining $\tilde{\mu}(\tilde{X}) := E[\tilde{M}_i \mid A_i = 1, \tilde{X}_i]$, we can re-write our functional of interest simply as $E_P[\tilde{\mu}(\tilde{X}_i)]$, i.e., it is also a MARS mean functional. Then, similarly to Proposition 5, the proof of Proposition 4 tells us that the nonparametric EIF is

$$\varphi_{\text{num}}(W_i) = \tilde{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(\tilde{X}_i)}(\tilde{M}_i - \tilde{\mu}(\tilde{X}_i)) - \theta_{\text{num}}.$$

Application of Lemma 1 then again delivers that this is also the EIF under $\mathcal{P}_\pi$. Under Assumption 2, and given that $[(Z_i, C_i) \perp\!\!\!\perp A_i] \mid X_i$, we recover that

$$\varphi_{\text{num}}(W_i) = \tilde{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(\tilde{M}_i - \tilde{\mu}(\tilde{X}_i)) - \theta_{\text{num}}.$$

We now want to prove that the efficient influence function for the numerator functional $E_P\left[Z_i^{\perp}Y_i\right]$ under $\mathcal{P}_\pi$ is

$$\varphi_{\text{num}}(W_i) = Z_i^{\perp}Y_i - \theta_{\text{num}}.$$

This follows immediately from the proof of the EIF for the denominator functional in Proposition 5, where $Z_i^{\perp}Y_i$ plays the role of ${C_{i,j}^{\perp}}^2$. □

9.1.8 Proposition 7

Proof. We first want to show that the efficient influence function for functional

$$E_P[\mu_C(\tilde{X}_i) \mid C_i = 1]$$

under the semiparametric statistical model $\mathcal{P}_\pi \ni P$ induced by Assumption 3 is

$$\varphi_C(W_i) = \frac{C_i}{P(C_i = 1)}\left[\mu_C(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu_C(\tilde{X}_i))\right] - \theta_C.$$

Note that, letting $\tilde{M}_i := \frac{C_i}{P(C_i=1)}M_i$ and $\tilde{M}_i^* := \frac{C_i}{P(C_i=1)}M_i^*$, we have that

$$E_P[\mu_C(\tilde{X}_i) \mid C_i = 1] = E_P[\tilde{\mu}(\tilde{X}_i, C_i)]$$

for $\tilde{\mu}(\tilde{X}_i, C_i) := E[\tilde{M}_i \mid A_i = 1, (\tilde{X}_i, C_i)]$, i.e., $E_P[\mu_C(X_i) \mid C_i = 1]$ is a MARS mean functional.

We then know from the proof of Proposition 4 that the nonparametric EIF for this functional is

$$\begin{aligned}
\varphi_C(W_i) &= \tilde{\mu}(\tilde{X}_i, C_i) + \frac{A_i}{\pi(\tilde{X}_i, C_i)}(\tilde{M}_i - \tilde{\mu}(\tilde{X}_i, C_i)) - \theta_C \\
&= \frac{C_i}{P(C_i = 1)}\left[\mu(\tilde{X}_i, C_i) + \frac{A_i}{\pi(\tilde{X}_i, C_i)}(M_i - \mu(\tilde{X}_i, C_i))\right] - \theta_C \\
&= \frac{C_i}{P(C_i = 1)}\left[\mu(\tilde{X}_i, C_i = 1) + \frac{A_i}{\pi(\tilde{X}_i, C_i = 1)}(M_i - \mu(\tilde{X}_i, C_i = 1))\right] - \theta_C \\
&= \frac{C_i}{P(C_i = 1)}\left[\mu_C(\tilde{X}_i) + \frac{A_i}{\pi(\tilde{X}_i, C_i = 1)}(M_i - \mu_C(\tilde{X}_i))\right] - \theta_C.
\end{aligned}$$

Because this functional is a MARS mean functional, we apply Lemma 1 to again conclude φ_C is also the semiparametric EIF under $\mathcal{P}_\pi$. Then, under Assumption 2, and given $[C_i \perp\!\!\!\perp A_i] \mid X_i$, we again have that

$$\varphi_C(W_i) = \frac{C_i}{P(C_i = 1)}\left[\mu_C(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu_C(\tilde{X}_i))\right] - \theta_C.$$

Given that θ_G has exactly the same form as θ_C, but with G_{i2} playing the role of C_i, the proof for the stated EIF for θ_G under $\mathcal{P}_\pi$ is immediate.

By the definition of pathwise differentiability, we have that

$$\begin{aligned}
\nabla_h \theta &= \nabla_h \theta_G - \nabla_h \theta_C \\
&= E[\varphi_G h] - E[\varphi_C h] \\
&= E[(\varphi_G - \varphi_C)h]
\end{aligned}$$

and so too that

$$\nabla_h = \nabla_{h_\pi} \theta = E[(\varphi_G - \varphi_C)h_\pi] = E[(\varphi_G - \varphi_C)h]$$

via Lemma 1. As such, we can conclude that the efficient influence function for θ under $\mathcal{P}_\pi \ni P$ is given by

$$\varphi := \varphi_G - \varphi_C.$$

□

9.1.9 Proposition 8

Proof. The proof of this proposition is immediate from Proposition 7, as the functionals have identical structure; the conditioning event $\{G_{i2} = 1\}$ is simply replaced by $\{R_i \in \mathcal{B}, D_i = 1\}$, and the conditioning event $\{C_i = 1\}$ is replaced by $\{R_i \in \mathcal{B}, D_i = 0\}$, and the proof follows identically. □

9.1.10 Theorem 1

Proof. To prove a CLT for $\hat{\theta}$ and show that it attains asymptotic efficiency, we will (1) introduce an "oracle" version of the MARS estimator where μ is known and show that this oracle estimator attains the semiparametric efficiency lower bound and then (2) show that the feasible MARS

estimator is asymptotically equivalent to the oracle estimator under our assumptions.

To start, recall we have the dataset $\{W_i\}_{i=1}^n$, $W_i \overset{\text{iid}}{\sim} P$. We partition the dataset randomly and disjointly into two parts, forming two new datasets: an "estimation sample" with indices given by the set $\mathcal{I}$, and a "training sample" with indices given by the set $\mathcal{I}'$. Equivalently, we may consider that we have access to two datasets, $\{W_i\}_{i\in\mathcal{I}}$ and $\{W_{i'}\}_{i'\in\mathcal{I}'}$, for which $W_i \overset{\text{iid}}{\sim} P$ and $W_{i'} \overset{\text{iid}}{\sim} P$.

Also recall that under Assumptions 1 and 2, we can point identify our functional of interest $\theta = E[M^*]$ as $\theta = E[\mu(\tilde{X})]$.

We now introduce the oracle estimator

$$\tilde{\theta} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i))\right].$$

Notice that we can re-write the oracle estimator as

$$\tilde{\theta} = \theta + \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\varphi(W_i)$$

for $\varphi(W_i)$ the EIF for θ under $\mathcal{P}_\pi$, as proved in Proposition 4. Rearranging and scaling, we have

$$\sqrt{|\mathcal{I}|}\left(\tilde{\theta} - \theta\right) = \frac{1}{\sqrt{|\mathcal{I}|}}\sum_{i\in\mathcal{I}}\varphi(W_i)$$

By application of Lindeberg-Lévy CLT for iid data, we have that

$$\sqrt{|\mathcal{I}|}\left(\tilde{\theta} - \theta\right) \underset{d}{\to} N(0, \text{Var}(\varphi(W_i))),$$

and as such we can see that $\tilde{\theta}$ is semiparametric efficient.

We now want to show that, under our assumptions,

$$\sqrt{|\mathcal{I}|}\left(\hat{\theta} - \tilde{\theta}\right) \underset{p}{\to} 0$$

where, as introduced earlier,

$$\hat{\theta} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i))\right]$$

where $\hat{\mu}$ is a random function of $\{W_{i'}\}_{i'\in\mathcal{I}'}$, the training sample. In particular, we will leverage Chebyshev's inequality, for which we need to show the mean of this scaled difference is zero, and its variance goes to zero asymptotically.

To start, we can write

$$\begin{aligned}\hat{\theta} - \tilde{\theta} &= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}\left(M_i - \hat{\mu}(\tilde{X}_i)\right) - \mu(\tilde{X}_i) - \frac{A_i}{\pi(X_i)}\left(M_i - \mu(\tilde{X}_i)\right)\right] \\ &= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\left(\hat{\mu}(\tilde{X}_i) - \mu(\tilde{X}_i)\right)\left(1 - \frac{A_i}{\pi(X_i)}\right)\right].\end{aligned}$$

Consider the expectation of this difference,

$$
\begin{aligned}
E\left[\hat{\theta}-\tilde{\theta}\right] &= E\left[\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right]\right] \\
&= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} E\left[\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right].
\end{aligned}
$$

Further considering an arbitrary summand (and abusing notation slightly, where conditioning on $\mathcal{I}'$ indicates conditioning on training dataset $\{W_{i'}\}_{i'\in\mathcal{I}'}$), we have that

$$
\begin{aligned}
E\left[\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right] &= E\left[E\left[\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right) \mid \tilde{X}_i, \mathcal{I}'\right]\right] \\
&= E\left[\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right) E\left[\left(1-\frac{A_i}{\pi(X_i)}\right) \mid \tilde{X}_i, \mathcal{I}'\right]\right] \\
&= E\left[\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{E\left[A_i \mid \tilde{X}_i, \mathcal{I}'\right]}{\pi(X_i)}\right)\right] \\
&= E\left[\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{E\left[A_i \mid X_i\right]}{\pi(X_i)}\right)\right] \\
&= 0,
\end{aligned}
$$

and as such it is immediate that $E\left[\hat{\theta}-\tilde{\theta}\right]=0$, and thus that $E\left[\sqrt{|\mathcal{I}|}\left(\hat{\theta}-\tilde{\theta}\right)\right]=0$. (N.B.: this holds for all $|\mathcal{I}|$.)

Now, computing the variance of this difference, we have

$$
\begin{aligned}
&\mathrm{Var}\left[\hat{\theta}-\tilde{\theta}\right] \\
&= \mathrm{Var}\left[\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right]\right] \\
&= \frac{1}{|\mathcal{I}|^2}\mathrm{Var}\left[\sum_{i\in\mathcal{I}}\left[\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right]\right] \\
&= \frac{1}{|\mathcal{I}|^2}E\left[\mathrm{Var}\left[\left(\sum_{i\in\mathcal{I}}\left[\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right]\right) \mid \tilde{X}_i, \mathcal{I}'\right]\right] \\
&= \frac{1}{|\mathcal{I}|}E\left[\mathrm{Var}\left[\left(\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right) \mid \tilde{X}_i, \mathcal{I}'\right]\right] \\
&= \frac{1}{|\mathcal{I}|}E\left[E\left[\left(\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right)^2 \mid \tilde{X}_i, \mathcal{I}'\right]\right] \\
&= \frac{1}{|\mathcal{I}|}E\left[\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)^2\frac{1-\pi(X_i)}{\pi(X_i)}\right] \\
&\leq \frac{1}{|\mathcal{I}|}\frac{1-\eta}{\eta}E\left[\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)^2\right] && \text{(Assumption 3)} \\
&= \frac{o(1)}{|\mathcal{I}|}. && \text{(Assumption 5)}
\end{aligned}
$$

As such, it is immediate that

$$\text{Var}\left[\sqrt{|\mathcal{I}|}\left(\hat{\theta}-\tilde{\theta}\right)\right] = |\mathcal{I}|\text{Var}\left[\left(\hat{\theta}-\tilde{\theta}\right)\right] = o(1).$$

Then, by Chebyshev's inequality, we have indeed shown that

$$\sqrt{|\mathcal{I}|}\left(\hat{\theta}-\tilde{\theta}\right) \underset{p}{\to} 0.$$

Accordingly, we have that

$$\sqrt{|\mathcal{I}|}\left(\hat{\theta}-\theta\right) = \sqrt{|\mathcal{I}|}\left(\tilde{\theta}-\theta\right) + o_p(1) \underset{d}{\to} N(0, \text{Var}(\varphi(W_i)))$$

by Slutsky's theorem. □

9.1.11 Theorem 2

Proof. In order to prove asymptotic efficiency for $\hat{\theta}_j$, we proceed with the following steps: (1) we prove CLTs for $\hat{\theta}_{j,\text{num}}$ and $\hat{\theta}_{j,\text{den}}$, and show that both are efficient estimators; (2) we prove a multivariate CLT for $\hat{\boldsymbol{\theta}}_{\boldsymbol{j}} = (\hat{\theta}_{j,\text{num}}, \hat{\theta}_{j,\text{den}})^\intercal$, and show that $\hat{\boldsymbol{\theta}}_{\boldsymbol{j}}$ is efficient; and (3) we appeal to van der Vaart (1998) Theorem 25.47, which states that the delta method preserves efficiency, to show that $\hat{\theta}_j$ is efficient.

Recall that

$$\hat{\theta}_{j,\text{num}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \check{C}_{i,j}\left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i))\right], \quad \hat{\theta}_{j,\text{den}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \check{C}_{i,j}^2.$$

We start by proving a CLT for $\hat{\theta}_{j,\text{den}}$. Note first and foremost that

$$\check{C}_{i,j} = C_{i,j}^{\perp} + o_p(1).$$

Then,

$$\begin{aligned}\hat{\theta}_{j,\text{den}} &= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}(C_{i,j}^{\perp} + o_p(1))^2 \\ &= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}({C_{i,j}^{\perp}}^2 + o_p(1))\end{aligned}$$

As such, we have that

$$\begin{aligned}
\sqrt{|\mathcal{I}|}\left(\hat{\theta}_{j,\text{den}} - \theta_{j,\text{den}}\right) &= \frac{1}{\sqrt{|\mathcal{I}|}}\sum_{i\in\mathcal{I}}\left({C_{i,j}^{\perp}}^2 - E[{C_{i,j}^{\perp}}^2] + o_p(1)\right) \\
&= \frac{1}{\sqrt{|\mathcal{I}|}}\sum_{i\in\mathcal{I}}\left({C_{i,j}^{\perp}}^2 - E[{C_{i,j}^{\perp}}^2]\right) + \frac{1}{\sqrt{|\mathcal{I}|}}\sum_{i\in\mathcal{I}} o_p(1) \\
&= \frac{1}{\sqrt{|\mathcal{I}|}}\sum_{i\in\mathcal{I}}\left({C_{i,j}^{\perp}}^2 - E[{C_{i,j}^{\perp}}^2]\right) + o(1)o_p(1) \\
&= \frac{1}{\sqrt{|\mathcal{I}|}}\sum_{i\in\mathcal{I}}\left({C_{i,j}^{\perp}}^2 - E[{C_{i,j}^{\perp}}^2]\right) + o_p(1) \\
&\underset{d}{\to} N(0, \text{Var}(\varphi_{\text{den}}(W_i))).
\end{aligned}$$

We now introduce the oracle estimator

$$\tilde{\theta}_{j,\text{num}} := \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} C_{i,j}^{\perp}\left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i))\right].$$

Using identical arguments as in Theorem 1, we know that this estimator is semiparametric efficient:

$$\sqrt{|\mathcal{I}|}\left(\tilde{\theta}_{j,\text{num}} - \theta_{j,\text{num}}\right) \underset{d}{\to} N(0, \text{Var}(\varphi_{\text{num}}(W_i))).$$

Now consider the "almost oracle" estimator

$$\begin{aligned}
\check{\theta}_{j,\text{num}} &:= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \check{C}_{i,j}\left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i))\right] \\
&= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}(C_{i,j}^{\perp} + o_p(1))\left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i))\right] \\
&= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} C_{i,j}^{\perp}\left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i))\right] + o_p(1)\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i))\right] \\
&= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} C_{i,j}^{\perp}\left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i))\right] + \frac{o_p(1)}{\sqrt{|\mathcal{I}|}}\frac{1}{\sqrt{|\mathcal{I}|}}\sum_{i\in\mathcal{I}}\left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i))\right] \\
&= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} C_{i,j}^{\perp}\left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i))\right] + \frac{o_p(1)}{\sqrt{|\mathcal{I}|}}O_p(1) \\
&= \tilde{\theta}_{j,\text{num}} + \frac{o_p(1)}{\sqrt{|\mathcal{I}|}}.
\end{aligned}$$

As such, it is clear that

$$\sqrt{|\mathcal{I}|}(\check{\theta}_{j,\text{num}} - \tilde{\theta}_{j,\text{num}}) = o_p(1),$$

i.e., $\check{\theta}_{j,\text{num}}$ is asymptotically equivalent to $\tilde{\theta}_{j,\text{num}}$.

Finally, then, we want to show that $\hat{\theta}_{j,\text{num}}$ is asymptotically equivalent to $\check{\theta}_{j,\text{num}}$. Notice

that

$$E\left[\hat{\theta}-\check{\theta}\right] = E\left[\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\check{C}_{i,j}\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right]\right]$$
$$= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} E\left[\check{C}_{i,j}\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right]$$

where

$$\begin{aligned}
E\left[\check{C}_{i,j}\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right] &= E\left[E\left[\check{C}_{i,j}\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right) \mid \tilde{X}_i, \mathcal{I}'\right]\right] \\
&= E\left[\check{C}_{i,j}\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right) E\left[\left(1-\frac{A_i}{\pi(X_i)}\right) \mid \tilde{X}_i, \mathcal{I}'\right]\right] \\
&= E\left[\check{C}_{i,j}\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{E\left[A_i \mid \tilde{X}_i, \mathcal{I}'\right]}{\pi(X_i)}\right)\right] \\
&= E\left[\check{C}_{i,j}\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{E\left[A_i \mid \tilde{X}_i\right]}{\pi(X_i)}\right)\right] \\
&= 0,
\end{aligned}$$

and so, just as in Theorem 1, $E\left[\sqrt{|\mathcal{I}|}(\hat{\theta}-\check{\theta})\right] = 0$.

Finally, observe that

$$\begin{aligned}
&\mathrm{Var}\left[\hat{\theta}-\check{\theta}\right] \\
&= \mathrm{Var}\left[\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\check{C}_{i,j}\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right]\right] \\
&= \frac{1}{|\mathcal{I}|^2}\mathrm{Var}\left[\sum_{i\in\mathcal{I}}\left[\check{C}_{i,j}\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right]\right] \\
&= \frac{1}{|\mathcal{I}|^2}E\left[\mathrm{Var}\left[\left(\sum_{i\in\mathcal{I}}\left[\check{C}_{i,j}\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right]\right) \mid \tilde{X}_i, \mathcal{I}'\right]\right] \\
&= \frac{1}{|\mathcal{I}|}E\left[\mathrm{Var}\left[\left(\check{C}_{i,j}\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right) \mid \tilde{X}_i, \mathcal{I}'\right]\right] \\
&= \frac{1}{|\mathcal{I}|}E\left[E\left[\left(\check{C}_{i,j}\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)\left(1-\frac{A_i}{\pi(X_i)}\right)\right)^2 \mid \tilde{X}_i, \mathcal{I}'\right]\right] \\
&= \frac{1}{|\mathcal{I}|}E\left[\check{C}_{i,j}^2\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)^2\frac{1-\pi(X_i)}{\pi(X_i)}\right] \\
&\leq \frac{1}{|\mathcal{I}|}\frac{1-\eta}{\eta}E\left[\check{C}_{i,j}^2\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)^2\right] && \text{(Assumption 3)} \\
&= \frac{1}{|\mathcal{I}|}\frac{1-\eta}{\eta}E\left[\check{C}_{i,j}^2 E\left[\left(\hat{\mu}(\tilde{X}_i)-\mu(\tilde{X}_i)\right)^2 \mid \check{C}_{i,j}\right]\right] \\
&= \frac{1}{|\mathcal{I}|}\frac{1-\eta}{\eta}E\left[\check{C}_{i,j}^2\right]o(1) && \text{(Assumption 5)} \\
&= \frac{o(1)}{|\mathcal{I}|},
\end{aligned}$$

where the second-to-last equality uses the fact that Assumption 5 holds conditionally (as this is directly implied by Assumption 5). As such, again by Chebyshev's inequality, identically to Theorem 1, we can conclude that

$$\sqrt{|\mathcal{I}|}(\hat{\theta}_{j,\text{num}} - \check{\theta}_{j,\text{num}}) = o_p(1)$$

indicating that $\hat{\theta}_{j,\text{num}}$ is asymptotically efficient, or, explicitly,

$$\sqrt{|\mathcal{I}|}\left(\hat{\theta}_{j,\text{num}} - \theta_{j,\text{num}}\right) \underset{d}{\to} N(0, \text{Var}(\varphi_{\text{num}}(W_i))).$$

We now form the vector-valued estimators and estimand

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{j}} := \begin{pmatrix} \hat{\theta}_{j,\text{num}} \\ \hat{\theta}_{j,\text{den}} \end{pmatrix}, \quad \tilde{\boldsymbol{\theta}}_{\boldsymbol{j}} := \begin{pmatrix} \tilde{\theta}_{j,\text{num}} \\ \tilde{\theta}_{j,\text{den}} \end{pmatrix}, \quad \boldsymbol{\theta}_{\boldsymbol{j}} := \begin{pmatrix} \theta_{j,\text{num}} \\ \theta_{j,\text{den}} \end{pmatrix}.$$

Notice then, by the (multivariate) CLT for iid data, leveraging our previous results,

$$\sqrt{|\mathcal{I}|}(\hat{\boldsymbol{\theta}}_{\boldsymbol{j}} - \boldsymbol{\theta}_{\boldsymbol{j}}) = \sqrt{|\mathcal{I}|}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{j}} - \boldsymbol{\theta}_{\boldsymbol{j}}) + o_p(1) \underset{d}{\to} N\left(0, E[\boldsymbol{\varphi}\boldsymbol{\varphi}^{\mathsf{T}}]\right)$$

where $\boldsymbol{\varphi} := \begin{pmatrix} \varphi_{\text{num}}(W_i) \\ \varphi_{\text{den}}(W_i) \end{pmatrix}$. In this form, we recognize that $\hat{\boldsymbol{\theta}}_{\boldsymbol{j}}$ is semiparametric efficient by van der Vaart (1998) Theorem 25.20.

Lastly, note that

$$\hat{\theta}_j = \phi(\hat{\boldsymbol{\theta}}_{\boldsymbol{j}}), \quad \theta_j = \phi(\boldsymbol{\theta}_{\boldsymbol{j}}),$$

for $\phi(a, b) = ab^{-1}$. Then, by application of the delta method, we have that

$$\sqrt{|\mathcal{I}|}(\hat{\theta}_j - \theta_j) \underset{d}{\to} N(0, \Sigma)$$

where, noting that,

$$\nabla\phi(a, b) = (b^{-1}, -ab^{-2})^{\mathsf{T}},$$

by the delta method we have that

$$\begin{aligned}
\Sigma &= \nabla\phi(\theta_{j,\text{num}}, \theta_{j,\text{den}})^{\mathsf{T}} E[\boldsymbol{\varphi}\boldsymbol{\varphi}^{\mathsf{T}}] \nabla\phi(\theta_{j,\text{num}}, \theta_{j,\text{den}}) \\
&= (\theta_{j,\text{den}}^{-1}, -\theta_{j,\text{num}}\theta_{j,\text{den}}^{-2}) \begin{pmatrix} \text{Var}(\varphi_{\text{num}}) & E[\varphi_{\text{num}}\varphi_{\text{den}}] \\ E[\varphi_{\text{num}}\varphi_{\text{den}}] & \text{Var}(\varphi_{\text{den}}) \end{pmatrix} \begin{pmatrix} \theta_{j,\text{den}}^{-1} \\ -\theta_{j,\text{num}}\theta_{j,\text{den}}^{-2} \end{pmatrix} \\
&= (\theta_{j,\text{den}}^{-1}, -\theta_{j,\text{num}}\theta_{j,\text{den}}^{-2}) \begin{pmatrix} \theta_{j,\text{den}}^{-1}\text{Var}(\varphi_{\text{num}}) - \theta_{j,\text{num}}\theta_{j,\text{den}}^{-2}E[\varphi_{\text{num}}\varphi_{\text{den}}] \\ \theta_{j,\text{den}}^{-1}E[\varphi_{\text{num}}\varphi_{\text{den}}] - \theta_{j,\text{num}}\theta_{j,\text{den}}^{-2}\text{Var}(\varphi_{\text{den}}) \end{pmatrix} \\
&= \theta_{j,\text{den}}^{-2}\text{Var}(\varphi_{\text{num}}) - 2\theta_{j,\text{num}}\theta_{j,\text{den}}^{-3}E[\varphi_{\text{num}}\varphi_{\text{den}}] + \theta_{j,\text{num}}^2\theta_{j,\text{den}}^{-4}\text{Var}(\varphi_{\text{den}}) \\
&= \theta_{j,\text{den}}^{-2}\left(\text{Var}(\varphi_{\text{num}}) - 2\theta_j E[\varphi_{\text{num}}\varphi_{\text{den}}] + \theta_j^2\text{Var}(\varphi_{\text{den}})\right) \\
&= \theta_{j,\text{den}}^{-2}\left(\text{Var}(\varphi_{\text{num}}) - 2E[\varphi_{\text{num}}\theta_j\varphi_{\text{den}}] + \text{Var}(\theta_j\varphi_{\text{den}})\right) \\
&= \theta_{j,\text{den}}^{-2}\text{Var}(\varphi_{\text{num}} - \theta_j\varphi_{\text{den}}) \\
&= \text{Var}(\theta_{j,\text{den}}^{-1}(\varphi_{\text{num}} - \theta_j\varphi_{\text{den}})) \\
&= \text{Var}(\varphi_j)
\end{aligned}$$

where

$$\varphi_j(W_i) = \theta_{j,\text{den}}^{-1}(\varphi_{\text{num}} - \theta_j\varphi_{\text{den}}) = \theta_{j,\text{den}}^{-1}C_{i,j}^{\perp}\left[\mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i)) - C_{i,j}^{\perp}\theta_j\right].$$

By van der Vaart (1998) Theorem 25.47, $\hat{\theta}_j$ is semiparametric efficient with efficient influence function φ_j. □

9.1.12 Theorem 3

Proof. The proof of the efficiency of $\hat{\theta}$ is immediate from the proof of Theorem 2. This can be seen by replacing $Z_i^{\perp}Y_i$ with ${C_{i,j}^{\perp}}^2$ in the proof of Theorem 2 (notice it is still the case that $\check{Z}_iY_i = Z_i^{\perp}Y_i + o_p(1)$ so long as we have that Y_i is stochastically bounded), as well as observing that the structure of (the estimator of) the numerator functional in Theorem 2 is the same as that of the denominator functional in the linear IV application, and vice versa. □

9.1.13 Theorem 4

Proof. In order to prove the efficiency of $\hat{\theta}$, we may proceed by: (1) proving the efficiency of $(\hat{\theta}_G, \hat{\theta}_C)$ for (θ_G, θ_C); and (2) applying the delta method to get a limiting distribution for $\hat{\theta}$, confirming that the asymptotic variance is the semiparametric lower bound $\text{Var}(\varphi(W_i))$.

First, we note that, just by the WLLN,

$$\hat{P}(G_{i2} = 1) = P(G_{i2} = 1) + o_p(1)$$

and so by the CMT,

$$\hat{P}(G_{i2} = 1)^{-1} = P(G_{i2} = 1)^{-1} + o_p(1),$$

and so overall we have that

$$G_{i2}\hat{P}(G_{i2} = 1)^{-1} = G_{i2}P(G_{i2} = 1)^{-1} + o_p(1).$$

With this observation, we now may use an identical proof for the efficiency of $\hat{\theta}_C$ as for the estimator of the numerator functional in Theorem 2, but where the role of $\check{C}_{i,j}$ is now played by $G_{i2}\hat{P}(G_{i2} = 1)^{-1}$, μ is played by μ_G, and $\tilde{X}_i$ is played by $(\tilde{X}_i, G_{i2})$. So too does the proof of the efficiency of $\hat{\theta}_C$ proceed identically, given that it has identical structure to $\hat{\theta}_G$, simply replacing conditioning event $\{G_{i2} = 1\}$ with $\{C_i = 1\}$ in the proof.

The output of application of these identically structured proofs yields that, for the vector-valued estimators and estimand

$$\hat{\boldsymbol{\theta}} := \begin{pmatrix}\hat{\theta}_G \\ \hat{\theta}_C\end{pmatrix}, \quad \tilde{\boldsymbol{\theta}} := \begin{pmatrix}\tilde{\theta}_G \\ \tilde{\theta}_C\end{pmatrix}, \quad \boldsymbol{\theta} := \begin{pmatrix}\theta_G \\ \theta_C\end{pmatrix},$$

(noting that $\tilde{\theta}_G, \tilde{\theta}_C$ are analogously defined oracle estimators), we can write

$$\sqrt{|\mathcal{I}|}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \sqrt{|\mathcal{I}|}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(1) \underset{d}{\to} N\left(0, E[\boldsymbol{\varphi}\boldsymbol{\varphi}^{\mathsf{T}}]\right)$$

where $\boldsymbol{\varphi} := \begin{pmatrix} \varphi_G \\ \varphi_C \end{pmatrix}$.

Lastly, note that

$$\hat{\theta} = \phi(\hat{\boldsymbol{\theta}}), \quad \theta = \phi(\boldsymbol{\theta}),$$

for $\phi(a, b) = a - b$, so, by application of the delta method, we have that

$$\sqrt{|\mathcal{I}|}(\hat{\theta} - \theta) \underset{d}{\to} N(0, \text{Var}(\varphi_G - \varphi_C)).$$

Because we know that $\varphi_G - \varphi_C := \varphi$ is the EIF for θ under $\mathcal{P}_\pi$, we know that $\text{Var}(\varphi_G - \varphi_C)$ is the semiparametric variance lower bound for θ, and as such $\hat{\theta}$ is efficient. (Alternatively, if we did not know φ was the EIF from Proposition 7, we could invoke van der Vaart (1998) Theorem 25.47 here to conclude that it must be.) □

9.1.14 Theorem 5

Proof. This theorem is immediate from the proof of Theorem 4, owing to their identical structure, but where, in the proof of Theorem 4: $G_{i2}\hat{P}(G_{i2} = 1)^{-1}$ is replaced by $\mathbb{I}\{R_i \in \mathcal{B}, D_i = 1\}\hat{P}(R_i \in \mathcal{B}, D_i = 1)^{-1}$ and $C_i\hat{P}(C_i = 1)^{-1}$ is replaced by $\mathbb{I}\{R_i \in \mathcal{B}, D_i = 0\}\hat{P}(R_i \in \mathcal{B}, D_i = 0)^{-1}$; and μ_G is replaced by μ_1 and μ_C is replaced by μ_0. □

9.2 A consistent estimator for the asymptotic variance of efficient estimators of MARS mean functionals

Consider the dataset $\{W_i\}_i^n$ with $W_i = (\tilde{M}_i, A_i, \tilde{X}_i)$. A MARS mean functional of interest is given by, under Assumptions 1 and 2,

$$\theta = E[M^*] = E[\tilde{\mu}(\tilde{X})]$$

for $\tilde{\mu}(\tilde{X}) = E[\tilde{M} \mid A = 1, \tilde{X}]$. Let $\check{\mu}(\tilde{X}) = \hat{E}[\tilde{M} \mid A = 1, \tilde{X}]$. By Theorem 1, we know that the estimator

$$\hat{\theta} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \hat{\varphi}_i,$$

with

$$\hat{\varphi}_i := \check{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(\tilde{M}_i - \check{\mu}(\tilde{X}_i))$$

is semiparametric efficient, with CLT

$$\sqrt{|\mathcal{I}|}\left(\hat{\theta} - \theta\right) \underset{d}{\to} N(0, \sigma^2)$$

where $\sigma^2 := \text{Var}(\varphi)$, and where

$$\varphi_i := \tilde{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(\tilde{M}_i - \tilde{\mu}(\tilde{X}_i)) - \theta$$

is the EIF for θ under $\mathcal{P}_\pi$, as follows from the result of Proposition 4.

It is then immediate from Slutksy's theorem that

$$\sigma^{-1}\sqrt{|\mathcal{I}|}\left(\hat{\theta}-\theta\right) \underset{d}{\to} N(0,1).$$

Of course, σ is not observed. However, so long as we have a consistent estimator $\hat{\sigma}$, we then have by the CMT and Slutksy again that

$$\hat{\sigma}^{-1}\sqrt{|\mathcal{I}|}\left(\hat{\theta}-\theta\right) \underset{d}{\to} N(0,1),$$

a result which can be used to form asymptotically valid confidence intervals based on critical values of the standard normal distribution.

In particular, for any MARS mean functional, consider the empirical variance estimator

$$\boxed{\hat{\sigma}^2 := \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\hat{\varphi}_i^2 - \hat{\theta}^2.}$$

Below, we prove that $\hat{\sigma}^2 = \sigma^2 + o_p(1)$, or that $\hat{\sigma}^2$ is a consistent variance estimator, allowing us to form confidence intervals using $\hat{\sigma}^2$ that hold in the large sample limit.

Proof. First, consider the oracle variance estimator

$$\tilde{\sigma}^2 := \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\varphi_i^2 - \hat{\theta}^2.$$

By the CMT, and from Proposition 4, we know that

$$\hat{\theta}^2 = \theta^2 + o_p(1).$$

As such, by the WLLN and Slutksy, we have that

$$\tilde{\sigma}^2 \underset{p}{\to} E[\varphi_i^2] - \theta^2 = \sigma^2.$$

Also recall from Theorem 1 that

$$\begin{aligned}
E[\hat{\theta}-\tilde{\theta}] &= 0 \\
E[\hat{\theta}-\tilde{\theta}] &= E\left[\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\hat{\varphi}_i\right] - E\left[\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\varphi_i\right] \\
&= E\left[\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}(\hat{\varphi}_i-\varphi_i)\right] \\
&= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}E\left[(\hat{\varphi}_i-\varphi_i)\right] \\
&= E\left[(\hat{\varphi}_i-\varphi_i)\right] \\
\therefore E\left[(\hat{\varphi}_i-\varphi_i)\right] &= 0
\end{aligned}$$

and that

$$\begin{aligned}
|\mathcal{I}|\text{Var}(\hat{\theta} - \tilde{\theta}) &= o(1) \\
|\mathcal{I}|\text{Var}(\hat{\theta} - \tilde{\theta}) &= |\mathcal{I}|\text{Var}\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\hat{\varphi}_i - \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\varphi_i\right) \\
&= |\mathcal{I}|\text{Var}\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}(\hat{\varphi}_i - \varphi_i)\right) \\
&= \frac{1}{|\mathcal{I}|}\text{Var}\left(\sum_{i\in\mathcal{I}}(\hat{\varphi}_i - \varphi_i)\right) \\
&= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\text{Var}\left(\hat{\varphi}_i - \varphi_i\right) \\
&= \text{Var}\left(\hat{\varphi}_i - \varphi_i\right) \\
\therefore \text{Var}\left(\hat{\varphi}_i - \varphi_i\right) &= o(1).
\end{aligned}$$

(N.B.: both of these conclusions hold because $\hat{\varphi}_i - \varphi_i \overset{\text{iid}}{\sim}$.) As a result, again by Chebyshev's inequality, we have that

$$\hat{\varphi}_i = \varphi_i + o_p(1)$$

and as such

$$\begin{aligned}
\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\hat{\varphi}_i^2 &= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}(\varphi_i + o_p(1))^2 \\
&= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\varphi_i^2 + o_p(1).
\end{aligned}$$

We can now conclude that

$$\hat{\sigma}^2 = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\hat{\varphi}_i^2 - \hat{\theta}^2 = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\varphi_i^2 - \hat{\theta}^2 + o_p(1) = \tilde{\sigma}^2 + o_p(1) = \sigma^2 + o_p(1).$$

□

9.3 Independent Measurement Error Models

We return to the setting of Section 6.3, but now allow measurement errors to be independent and *not* identically distributed:

$$X_{k,i} = X_{k,i}^* + \eta_{k,i}, \quad \eta_{k,i} \overset{d}{\approx} N(0, \sigma_{\eta,k,i}^2), \quad \sigma_{\eta,k,i}^2 := |\mathcal{I}_{k,i}|^{-1}\text{Var}(\varphi_{k,i}),$$

where $\varphi_{k,i}$ is the efficient influence function associated with $X_{k,i}^*$.

We therefore will assume that

$$Y_i \perp\!\!\!\perp \eta_i, \quad \eta_i \overset{\text{ind}}{\sim} N(0, \Sigma_i), \quad \Sigma_i := \text{diag}(\sigma_{\eta,1,i}^2, \ldots, \sigma_{\eta,K,i}^2),$$

owing to the facts that (a) it will often be the case that each $X_{k,i}$ will often be estimated on a separate sample, (b) for large $\mathcal{I}_{k,i}$ the normal approximation should be a reasonable one. We

assume Σ_i are known.

Let $\bar{\Sigma} := \frac{1}{n}\sum_{i=1}^n \Sigma_i$, and consider the estimator

$$\hat{\beta} := (X^\intercal X - n\bar{\Sigma})^{-1} X^\intercal Y.$$

This estimator is consistent, as we can see that

$$\begin{aligned}
\hat{\beta} &:= \left(X^\intercal X - n\bar{\Sigma}\right)^{-1} X^\intercal Y \\
&= \left(\frac{1}{n} X^\intercal X - \bar{\Sigma}\right)^{-1} \frac{1}{n} X^\intercal Y \\
&= \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\intercal - \bar{\Sigma}\right)^{-1} \frac{1}{n}\sum_{i=1}^n X_i Y_i \\
&= \left(\frac{1}{n}\sum_{i=1}^n X_i^* X_i^{*\intercal} + \frac{1}{n}\sum_{i=1}^n \eta_i X_i^{*\intercal} + \frac{1}{n}\sum_{i=1}^n X_i^* \eta_i^\intercal + \frac{1}{n}\sum_{i=1}^n \eta_i \eta_i^\intercal - \bar{\Sigma}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n X_i^* Y_i + \frac{1}{n}\sum_{i=1}^n \eta_i Y_i\right) \\
&\underset{p}{\to} E[X_i^* X_i^{*\intercal}] E[X_i^* Y_i] = \beta.
\end{aligned}$$

where the last line follows by the WLLN, the CMT, and Kolmogorov's variance criterion for averages (c.f., corollary 5.22, Kallenberg (2021)), which yields

$$\frac{1}{n}\sum_{i=1}^n \eta_i \eta_i^\intercal - \bar{\Sigma} = \frac{1}{n}\sum_{i=1}^n (\eta_i \eta_i^\intercal - \Sigma_i) \underset{p}{\to} 0.$$

In order to understand the asymptotic variance of this estimator, note that

$$\begin{aligned}
\hat{\beta} &= \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^\intercal - \bar{\Sigma}\right)^{-1} \frac{1}{n}\sum_{i=1}^n X_i Y_i \\
&= \left(\frac{1}{n}\sum_{i=1}^n (X_i X_i^\intercal - E[X_i X_i^\intercal]) + E[X_i^* X_i^{*\intercal}]\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n (X_i Y_i - E[X_i Y_i]) + E[X_i^* Y_i]\right) \\
&= (\Delta_1 + \Omega)^{-1} (\Delta_2 + S),
\end{aligned}$$

where $\Delta_1 := \frac{1}{n}\sum_{i=1}^n (X_i X_i^\intercal - E[X_i X_i^\intercal])$, $\Delta_2 := \frac{1}{n}\sum_{i=1}^n (X_i Y_i - E[X_i Y_i])$, $\Omega := E[X_i^* X_i^{*\intercal}]$, and $S := E[X_i^* Y]$. Thus, in this notation, $\beta = \Omega^{-1} S$. Noting that both $\Delta_1 = O_p(n^{1/2})$ and $\Delta_2 = O_p(n^{1/2})$ by the Lindeberg-Feller CLT, by a Taylor expansion argument, we have that

$$\begin{aligned}
\hat{\beta} &= (\Delta_1 + \Omega)^{-1} (\Delta_2 + S) \\
&= (\Omega^{-1} - \Omega^{-1}\Delta_1\Omega^{-1} + O_p(n^{-1}))(\Delta_2 + S) \\
&= \Omega^{-1}\Delta_2 + \Omega^{-1} S - \Omega^{-1}\Delta_1\Omega^{-1} S + O_p(n^{-1}) \\
&= \beta + \Omega^{-1}\Delta_2 - \Omega^{-1}\Delta_1\beta + O_p(n^{-1})
\end{aligned}$$

and thus

$$\hat{\beta} - \beta = \Omega^{-1}(\Delta_2 - \Delta_1\beta) + O_p(n^{-1}).$$

Substituting back in and scaling, we have that

$$\sqrt{n}\left(\hat{\beta}-\beta\right) = \Omega^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i - E[Z_i]) + o_p(1),$$

where $Z_i := X_i(Y_i - X_i^\intercal\beta) \overset{\text{ind}}{\sim}$. Then by the Lindeberg-Feller CLT,

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \underset{d}{\to} N(0, \Omega^{-1}\Psi\Omega^{-1}).$$

where $\Psi := \lim_{n\to\infty}\frac{1}{n}\sum_i \mathrm{Var}(Z_i)$, assuming this limit exists.

Let $\hat{\Omega} := \frac{1}{n}\sum_i(X_iX_i^\intercal - \Sigma_i)$ and let $\hat{\Psi} := \frac{1}{n}\sum_i \hat{Z}_i\hat{Z}_i^\intercal - \bar{Z}_i\bar{Z}_i^\intercal$, for $\bar{Z}_i := \frac{1}{n}\sum_i \hat{Z}_i$ and $\hat{Z}_i = X_i(Y_i - X_i^\intercal\hat{\beta})$. Observe that, again by Kolmogorov's variance criterion for averages,

$$\hat{\Omega} \underset{p}{\to} \Omega, \quad \hat{\Psi} \underset{p}{\to} \Psi.$$

Then,

$$\hat{V}^{-1/2}\sqrt{n}\left(\hat{\beta}-\beta\right) \underset{d}{\to} N(0, I),$$

where $\hat{V} := \hat{\Omega}^{-1}\hat{\Psi}\hat{\Omega}^{-1}$.