INFERENCE FOR REGRESSION WITH VARIABLES GENERATED BY AI OR MACHINE LEARNING

By

Laura Battaglia, Timothy Christensen,

Stephen Hansen and Szymon Sacher

December 2024

COWLES FOUNDATION DISCUSSION PAPER NO. 2421

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Inference for Regression with Variables Generated by AI or Machine Learning*

Laura Battaglia
Oxford

Timothy Christensen
Yale

Stephen Hansen
UCL, IFS, and CEPR

Szymon Sacher
Meta

December 9, 2024

Abstract

It has become common practice for researchers to use AI-powered information retrieval algorithms or other machine learning methods to estimate variables of economic interest, then use these estimates as covariates in a regression model. We show both theoretically and empirically that naively treating AI- and ML-generated variables as "data" leads to biased estimates and invalid inference. We propose two methods to correct bias and perform valid inference: (i) an explicit bias correction with bias-corrected confidence intervals, and (ii) joint maximum likelihood estimation of the regression model and the variables of interest. Through several applications, we demonstrate that the common approach generates substantial bias, while both corrections perform well.

JEL Codes: C11, C51, C55

Keywords: Measurement Error, Artificial Intelligence, Large Language Models, Topic Models, Inference

*Authors are listed in alphabetical order. SH acknowledges funding from ERC Consolidator Grant 864863, which supported his and LB's time. We thank Nick Bloom, Germain Gauthier, Evan Munro, David Rossell, and Leif Thorsrud for feedback, as well as seminar and workshop participants at Aarhus, Bocconi, BSE, Bates, Columbia, ETH Zurich, LSE, Kent, Reserve Bank of Australia, UCSD, UPenn, USC, Warwick, the 3rd Monash-Warwick-Zurich Text-as-Data Workshop, the 2024 BSE Summer Institute, the 2024 FinEML Conference (USI Lugano), the 2024 Machine Learning in Economics Summer Conference (UChicago), the 2024 NASM (Vanderbilt), ESIF-AIML (Cornell), and ESAM (Monash) Conferences of the Econometric Society, the 2024 International Symposium on Nonparametric Statistics, and the 2024 ECONDAT Fall Meeting. We also thank Kirill Safonov for excellent research assistance.

1 Introduction

Unstructured and high-dimensional data is increasingly used to estimate latent variables of economic interest. Early and influential work includes Baker et al. (2016) which measures economic policy uncertainty from newspaper text and Hoberg and Phillips (2016) which groups firms into industries based on corporate filings. The advent of sophisticated machine learning (ML) and artificial intelligence (AI) algorithms has led to a wave of new research. For example, Magnolfi et al. (2022) use survey data to measure product differentiation; Compiani et al. (2023) measure substitutability between products using Amazon text and image data; Gorodnichenko et al. (2023) measure tone-of-voice from audio recordings of FOMC press conferences; Gabaix et al. (2023) impute firm characteristics from investor holdings data; Einav et al. (2022) infer patients' health status from surveys; Vafa et al. (2023) measure labor market experience from CVs.

These AI- and ML-derived measures are rarely an end in themselves. Rather, the goal is to study how the variables they proxy interact with the economic environment. In common practice, these estimated measures are plugged-in to downstream econometric models whose parameters are the main object of study. The *upstream* information retrieval (IR) model used to extract measurements from unstructured or high-dimensional data and the *downstream* econometric model are almost always taken as wholly separate: the output of the upstream model is treated as regular numeric "data" when estimating and performing inference on the downstream model. We call this the *two-step strategy*.

While clearly a pragmatic initial approach, the two-step strategy has largely unknown statistical properties. One intuition is that point estimation in the downstream model is biased due to measurement error in the upstream model. Another is that inference suffers from a generated regressor problem (Pagan 1984). At the same time, results in the time-series literature suggest plugging-in estimated latent variables need not lead to inference problems (Stock and Watson 2002, Bernanke et al. 2005, Bai and Ng 2006). Without a coherent framework for analyzing the problem, it is difficult to assess which of these perspectives is correct. More generally, characterizing the statistical guarantees—or lack thereof—of the two-step strategy is an important step in establishing a more mature understanding of reliable inference methods when working with variables generated by AI or ML, an area that is still in its infancy.

To study this problem, we consider a downstream regression in which an outcome (Y_i) depends on latent variables ($\boldsymbol{\theta}_i$) and observed variables ($\mathbf{q}_i$), with n total data points. The researcher has available an unstructured dataset ($\mathbf{x}_i$) for estimating $\boldsymbol{\theta}_i$.[1] The widely-used two-step strategy first computes an estimate $\hat{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ for each observation, then regresses

[1] We stay agnostic on the form that $\mathbf{x}_i$ takes. In the text setting, $\mathbf{x}_i$ is a sequence of words; in the image setting, it is matrices of RGB values; in the audio settings, it is a sequence of sampled amplitudes.

Y_i on $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{q}_i$ and reports point estimates and confidence intervals using standard OLS methods (i.e., treating $\hat{\boldsymbol{\theta}}_i$ as regular numeric data). Depending on the setting, one may wish to do inference on the coefficients on the latent or observed variables. In either case, the key question is whether this two-step approach leads to valid inference.

This environment can accommodate numerous different interpretations. One is *label imputation* in which $\boldsymbol{\theta}_i$ is a vector of binary labels, for example race or gender indicators. A researcher can use $\mathbf{x}_i$, for example photographs, as inputs into a classifier which outputs predicted label values. Another is *dimensionality reduction* in which the researcher estimates a low-dimensional representation of $\mathbf{x}_i$ using an unsupervised learning model. Yet another is *index construction from unstructured data*. For example, $\mathbf{x}_i$ could be a set of newspaper articles which the researcher individually classifies as containing positive or negative sentiment. These individual labels can then be aggregated and normalized to produce a sentiment value for each observation.

The measurement error associated with observation i is $\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i$, which can arise from various sources depending on the setting. In label imputation, classification error introduces a positive probability that $\hat{\boldsymbol{\theta}}_i \neq \boldsymbol{\theta}_i$.[2] In unsupervised learning, $\mathbf{x}_i$ is a noisy realization from a distribution that depends on $\boldsymbol{\theta}_i$, meaning $\hat{\boldsymbol{\theta}}_i$ is a noisy estimate of $\boldsymbol{\theta}_i$. Valid inference must account for both measurement error in the upstream model and sampling error in the downstream model. To ensure both forces are present, we adopt an asymptotic framework in which the precision of $\hat{\boldsymbol{\theta}}_i$ increases with n such that, as $n \to \infty$, measurement error in the upstream model is comparable to sampling error in the downstream regression.[3] In this way, our asymptotic analysis closely mirrors the finite-sample problem faced in practice. It also captures the prevailing trend whereby increasingly large datasets are analyzed by increasingly accurate algorithms.

In this framework, we derive two important results. First, the asymptotic distribution of the OLS estimator has a first-order bias due to measurement error in the upstream model. The magnitude of the bias is increasing in the size of the measurement error relative to sampling uncertainty in the downstream model. Second, the asymptotic variance of the OLS estimator is the same as if Y_i were regressed on the true $\boldsymbol{\theta}_i$ and $\mathbf{q}_i$. Moreover, OLS standard errors from the two-step strategy are consistent. As a result, two-step confidence intervals have the correct width but incorrect centering, leading to under-coverage and making them invalid for inference. This differs from a generated regressor problem, where the asymptotic variance is inflated but there is no location shift. To the extent that the literature acknowledges the two-step strategy might be a problem, concerns typically

[2] Algorithms are sometimes fine-tuned on small validation datasets. While this can reduce misclassification error, it introduces additional statistical dependencies that further complicate inference.

[3] Our use of sequences of DGPs to better approximate the finite-sample behavior of estimators is similar in spirit to the weak instrument literature (Staiger and Stock 1997), earlier work on measurement error (Chesher 1991), unit root testing (Phillips 1987), and large n, T panels (Hahn and Kuersteiner 2002).

focus on standard errors. Our analysis shows these concerns are misplaced: the primary issue is bias, not incorrect standard errors.

Having established the invalidity of the two-step strategy, we next propose two methods to correct bias and perform valid inference. The first is an explicit bias correction with bias-corrected confidence intervals. Because we characterize the first-order bias analytically, one can construct bias-corrected estimates and confidence intervals whenever the first-order bias is consistently estimable. We provide consistent estimators for the cases of label imputation and dimensionality reduction, which encompass many applications.

For label imputation in particular, bias correction requires estimating the false-positive rate from a validation sample of size m in which both $\boldsymbol{\theta}_i$ and $\hat{\boldsymbol{\theta}}_i$ are observed. A recent literature (Fong and Tyler 2021, Allon et al. 2023, Egami et al. 2023, Zhang et al. 2023) has proposed using validation data to implement an IV strategy for inference with AI- or ML-generated variables. IV strategies typically entail estimating a first-stage regression of $\boldsymbol{\theta}_i$ on $\hat{\boldsymbol{\theta}}_i$ in the validation sample, then regressing Y_i on the fitted values in the full dataset (of size n). The asymptotic environment that justifies IV-based inference assumes that m and n are of comparable size, as in the classical literature on auxiliary data (see Chen et al. 2008, for example). In contrast, our bias correction is valid even when n grows faster than m. This distinction is important because there is typically vastly more data than humans can reliably code by hand. Using a small simulation, we provide evidence that our bias-corrected confidence intervals perform well when n is much larger than m, whereas IV-based confidence intervals under-cover.

While convenient to apply, bias-correction formulas are not immediately available for all upstream specifications. As such, our second solution is to jointly estimate the latent variables and regression model by maximum likelihood. We refer to this as the *one-step strategy*. While implementing the one-step strategy is straightforward theoretically, it presents a major computational challenge due to the large number of latent variables. To address this, we use Hamiltonian Monte Carlo (HMC; MacKay 2003, Neal 2012), a Markov Chain Monte Carlo algorithm that uses information on the gradient of a distribution to sample from it. Implementation is greatly simplified with the use of modern probabilistic programming languages: one simply specifies the likelihood in code, which is then "automatically" compiled to perform sampling.[4]

We also assess the performance of the bias correction and one-step strategies in applied settings. The first application is label imputation. Simulation evidence shows that measurement error from misclassification biases two-step estimation and inference, whereas our bias-corrected estimates and confidence intervals perform well. We then explore the

[4]Previous papers that have performed inference using the joint likelihood approach with unstructured data include Gentzkow et al. (2019), Ruiz et al. (2020), and Munro and Ng (2022). These typically require extensive code to estimate, which makes adapting the model difficult for non-specialists.

dataset from Hansen et al. (2023), which generates a binary label indicating remote work for each job ad in the Lightcast dataset of online job postings. The classifier, based on a large language model trained against thousands of human labels, has high accuracy. Two-step regressions of posting-level wages on remote work indicators by occupation point to a positive association. However, most two-step estimates fall *outside* the bias-corrected confidence intervals, showing that the two-step strategy can severely distort inference.

Next, we analyze a setting where a dimensionality reduction algorithm generates the regressor. To this end, we introduce the Supervised Topic Model with Covariates (STMC) which combines elements of existing models (Blei et al. 2003, Roberts et al. 2014, Ahrens et al. 2021) but is, to the best of our knowledge, a new statistical model of unstructured data. The model reduces the dimensionality of feature-count vectors by projecting them onto a set of latent factors (or topics), as in Probabilistic Latent Semantic Analysis (Hofmann 1999) and Latent Dirichlet Allocation (Blei et al. 2003). The dependence of outcome variables on latent factor loadings and observed covariates is captured by a downstream regression model.

In simulated data, we show that the bias of two-step estimates is increasing in the relative importance of measurement error. Moreover, two-step confidence interval widths are similar to those obtained using the true latent variables as covariates. Both of these findings reinforce the main predictions of our theory. By contrast, the one-step strategy produces estimates that appear unbiased and confidence intervals that have both the correct width and the correct centering.

Finally, we revisit the empirical application from Bandiera et al. (2020) which uses the two-step strategy to first estimate latent CEO behaviors from a CEO time-use survey, then explains firm performance using the estimated behaviors. To explore the effect of increasing measurement error, we repeat the analysis sampling only 10% of the available information per CEO. Coherent with the predictions of our theory, we observe that the difference between the one-step and two-step estimates is much more notable with this reduced sample but the widths of confidence intervals are comparable in both cases.

Our overall message is that the increasingly common practice of using regressors generated by AI or ML introduces measurement error which biases estimates and leads to invalid inference. We illustrate this formally with theoretical arguments in specific, empirically relevant settings, but the take-away applies much more broadly. On a more positive note, solutions exist that are practical to implement. We therefore see bias correction and the one-step strategy as robust and widely applicable starting points for empirical analysis. For instance, an emerging line of research uses text-derived sentiment indices as inputs into forecasting models with a vector autoregressive or dynamic factor structure. Straightforward extensions of our theoretical arguments can be used to show how error in the indices will bias coefficient estimates and limit the effectiveness of these forecasting

methods. More constructively, our solutions can be used to enhance the performance of these forecasting methods. Likewise, the industrial organization literature is increasingly using embedded representations of firms and products to characterize market behavior and demand with structural models. Our solutions can be used to mitigate bias introduced by measurement error in the embeddings in these. Going forward, it is important to establish for which specific IR methods and econometric models does measurement error most severely affect inference. More generally, our belief is that inference problems arising from using the outputs of AI/ML algorithms should be better recognized and taken more seriously in order to fully harness their potential value.

The rest of the paper proceeds as follows. Section 2 provides a simple setting in which the inference problems associated with the two-step strategy emerge. Section 3 further develops these arguments and presents our main theoretical results, which we specialize to imputed labels and dimensionality reduction in Appendix B. Section 4 presents the bias correction and one-step strategies for valid inference. Section 5 provides empirical illustrations. Section 6 concludes.

2 Motivating Example

This section presents a stylized model showing how the two-step strategy leads to biased inference.

2.1 Stylized Model

The model is loosely based on Baker et al. (2016), which develops text-based measures of economic policy uncertainty (EPU) and explores the relationship between EPU and economic outcomes. Suppose we are interested in the effect γ_1 of θ_i (policy uncertainty in month i) on Y_i (employment or investment, say, in month $i+1$) in the regression model

$$Y_i = \gamma_0 + \gamma_1 \theta_i + \varepsilon_i. \tag{1}$$

Policy uncertainty is a nebulous concept that is difficult to precisely define let alone observe. Baker et al. (2016) form EPU indices from monthly counts of articles in ten newspapers containing certain terms, which they convert to an index. Their EPU index is used as a covariate in regressions and VARs. While the index is a strong signal of policy uncertainty, is not numerically the same as policy uncertainty. For instance, one could change the set of newspapers surveyed and obtain a quantitatively different (but related) measure. We therefore work with the model

$$X_i \sim \text{Binomial}(C_i, \theta_i), \tag{2}$$

where X_i is the number of articles containing specific terms in month i, C_i is the total number of articles that month, and θ_i is the policy uncertainty rate. The variables X_i, Y_i, and C_i are observed but θ_i is not. One can estimate θ_i using $\hat{\theta}_i = X_i/C_i$, as done by Baker et al. (2016, p. 1599) to construct their policy uncertainty measure.

To facilitate the theoretical derivations below, let $\mathbb{E}[\varepsilon_i|\theta_i, X_i, C_i] = 0$ and $\text{Var}(\theta_i) > 0$, so the OLS estimator of γ_1 would be consistent if θ_i were observed, and $\mathbb{E}[\varepsilon_i^2] < \infty$. We also assume (i) Y_i and (X_i, C_i) are independent conditional on θ_i, and (ii) C_i and θ_i are independent. These assumptions, which are credible in the context of Baker et al. (2016), are made primarily for convenience and can be relaxed. We assume the data are a random sample $(X_i, Y_i, C_i)_{i=1}^n$. Our analysis and findings extend to time-series data, though we focus on the IID case to simplify presentation.

2.2 Two-Step Strategy

In this example, the typical two-step strategy regresses Y_i on $\hat{\theta}_i$ and uses standard OLS inference for γ_1. This approach ignores the fact that $\hat{\theta}_i$ is a noisy estimate of θ_i, potentially leading to biased estimates and inference.

Let $\hat{\gamma}_1$ denote the OLS estimator of γ_1 from regressing of Y_i on $\hat{\theta}_i$. By standard OLS algebra, as the sample size $n \to \infty$ we have

$$\begin{aligned}\hat{\gamma}_1 &\to_p \gamma_1 \frac{\text{Cov}(\theta_i, \hat{\theta}_i)}{\text{Var}(\hat{\theta}_i)} \\ &= \gamma_1 \frac{\text{Var}(\theta_i)}{\text{Var}(\theta_i) + \mathbb{E}\left[C_i^{-1}\right] \mathbb{E}\left[\theta_i(1-\theta_i)\right]},\end{aligned}$$

because $\mathbb{E}[\hat{\theta}_i|\theta_i, C_i] = \theta_i$ and $\text{Var}(\hat{\theta}_i) = \text{Var}(\theta_i) + \mathbb{E}[C_i^{-1}]\mathbb{E}[\theta_i(1-\theta_i)]$ by the law of total variance and independence of C_i and θ_i. Evidently, there is an attenuation bias caused by measurement error in $\hat{\theta}_i$ which makes $\hat{\gamma}_1$ inconsistent.

The key determinant of bias is the average reciprocal amount of unstructured data per observation, $\mathbb{E}[C_i^{-1}]$. If the amount of unstructured data per observation is large, making $\mathbb{E}[C_i^{-1}]$ small, we have

$$\text{plim}(\hat{\gamma}_1) \approx \gamma_1 - \mathbb{E}\left[\frac{1}{C_i}\right] \frac{\mathbb{E}\left[\theta_i(1-\theta_i)\right]}{\text{Var}(\theta_i)} \gamma_1$$

because $(1+x)^{-1} \approx 1 - x$ for small x. Hence, the bias is of the order of $\mathbb{E}[C_i^{-1}]$.

In practice, both measurement error and sampling error can play important roles. To capture this, we consider a sequence of populations indexed by the sample size n. The distribution of (Y_i, X_i, θ_i) conditional on C_i is fixed but the distribution of C_i is changing

with n so that

$$\sqrt{n} \times \mathbb{E}\left[\frac{1}{C_i}\right] \to \kappa \in [0, \infty). \tag{3}$$

Working with this sequence of DGPs rather than a single fixed DGP allows us to gain insights about how $\hat{\gamma}_1$ behaves when both measurement and sampling error are present, as is the case in finite samples. The parameter κ controls the relative importance of measurement error: $\kappa = 0$ means sampling error dominates, while positive κ means both are of the same order. In that case, larger values of κ give relatively greater importance to measurement error.

Proposition 1. *Consider the sequence of populations just described. Then*

$$\sqrt{n}(\hat{\gamma}_1 - \gamma_1) \to_d N\left(-\kappa\,\gamma_1 \frac{\mathbb{E}[\theta_i(1-\theta_i)]}{\mathrm{Var}(\theta_i)}, \frac{\mathbb{E}[\varepsilon_i^2(\theta_i - \mathbb{E}[\theta_i])^2]}{\mathrm{Var}(\theta_i)^2}\right).$$

Proposition 1 shows that two-step inference is *valid* when $\kappa = 0$ because measurement error vanishes faster than sampling error. In this case, the estimated $\hat{\theta}_i$ can be treated as if they are the true θ_i. However, two-step inference is *invalid* when $\kappa > 0$. In this case, $\hat{\gamma}_1$ is consistent and its asymptotic variance is the same as if Y_i were regressed on the true θ_i, but the center of the asymptotic distribution is shifted due to measurement error. Confidence intervals based on standard OLS inference will therefore have approximately correct width but incorrect centering, so their coverage rates will be below nominal coverage.

3 Full Analysis of the Two-Step Strategy

In this section, we first describe the statistical framework linking unstructured data and the downstream regression model. We then analyze the two-step strategy and show why it leads to biased inference in empirically plausible settings.

3.1 Model

We begin by specifying a statistical model that, broadly speaking, has two parts. The first computes low-dimensional numerical representations of the unstructured data to proxy some latent variable of economic interest (e.g., "policy uncertainty" or "sentiment"). The second introduces these representations as covariates, along with other quantitative data, into a linear regression model. Formally, consider

$$Y_i = \boldsymbol{\gamma}^T \boldsymbol{\theta}_i + \boldsymbol{\alpha}^T \mathbf{q}_i + \varepsilon_i, \qquad \mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{\theta}_i, \mathbf{q}_i\right] = 0, \tag{4}$$

where $\boldsymbol{\theta}_i$ is a vector of latent variables of economic interest and $\mathbf{q}_i$ is a vector of observed quantitative data. For each observation i we observe unstructured data $\mathbf{x}_i$ from which

an estimate $\hat{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ is derived using an information retrieval model. The parameter $\boldsymbol{\gamma}$ is the key object of interest in most applications. But in some cases (e.g., Avivi (2024)), $\boldsymbol{\alpha}$ is the focus and $\boldsymbol{\theta}_i$ serves as a control variable derived from unstructured data.

The dominant two-step strategy can be summarized as follows:

(i) Estimates $\hat{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ are computed from unstructured data $\mathbf{x}_i$ using an upstream information retrieval model.

(ii) Y_i is regressed on $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{q}_i$. Inference is performed treating the $\hat{\boldsymbol{\theta}}_i$ as if they are regular numeric data.

Evidently there is a measurement error problem: the estimates $\hat{\boldsymbol{\theta}}_i$ are noisy proxies for the true $\boldsymbol{\theta}_i$ in the regression model (4). Step (ii) overlooks this problem and treats the estimates $\hat{\boldsymbol{\theta}}_i$ as regular numeric data. This raises the possibility of bias in the OLS estimates due to measurement error introduced in Step (i). Moreover, conventional standard errors are typically reported. But these do not account for the additional variation from using $\hat{\boldsymbol{\theta}}_i$ instead of $\boldsymbol{\theta}_i$, raising the possibility of a generated regressors problem. To understand the forces at play, we analyze the two-step strategy and formally demonstrate why it can lead to biased estimates and inference.

We first highlight some key applications that will serve as running examples. These applications are not exhaustive, and many others will exhibit similar behavior when used in the two-step strategy.

3.1.1 Application 1: AI/ML-Generated Labels.

It is increasingly common to use AI or ML methods to impute missing covariates from unstructured data. A leading use case involves regressions of an outcome Y_i on a latent binary variable θ_i (e.g., the sentiment of a news article or racial group membership) and observed controls $\mathbf{q}_i$.[5] Unstructured data $\mathbf{x}_i$ (e.g., article text or voter registration data) is often used to predict θ_i using a pre-trained classification algorithm. The two-step strategy entails first generating a prediction $\hat{\theta}_i$ of θ_i then regressing Y_i on $\hat{\theta}_i$ and $\mathbf{q}_i$. Here the source of measurement error is misclassification: $\hat{\theta}_i$ may differ from θ_i for some observations. It is important to note that high-quality classifiers deployed on large data sets may have misclassification rates that are non-negligible relative to sampling error.

3.1.2 Application 2: Topic Models.

A large empirical literature uses topic models as part of the two-step strategy. Examples include Hansen et al. (2018), Mueller and Rauh (2018), Larsen and Thorsrud (2019), Thorsrud (2020), Adams et al. (2021), Ash et al. (2022), and Bybee et al. (2024) with

[5] We present the case of scalar $\boldsymbol{\theta}_i$ here and defer the case of multiple categories to Appendix B.

text data, Bandiera et al. (2020), Draca and Schwarz (2021), and Munro and Ng (2022) with survey data, and Nimczik (2017) and Olivella et al. (2021) with network data.

Each unstructured observation i is represented by $\mathbf{x}_i = (x_{i,v})_{v=1}^V$, a V-dimensional vector of counts. The value $x_{i,v}$ is the number of times v appears in observation i. In the bag-of-words model, V is the number of unique terms in a textual corpus—typically in the thousands—and $x_{i,v}$ is the count of term v in document i. Since $\mathbf{x}_i$ is a count vector, it can be modeled as a Multinomial without loss of generality. For interpretability, a factor structure is imposed on the count probabilities, as in Probabilistic Latent Semantic Analysis (Hofmann 1999) and Latent Dirichlet Allocation (Blei et al. 2003, LDA).

There are $K \ll V$ separate distributions over the V features, denoted $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, where each $\boldsymbol{\beta}_k \in \Delta^{V-1}$, the $(V-1)$-dimensional simplex. The distributions are called *topics* in text applications. More generally, they represent common factors from which individual observations are built. We collect the factors into a $K \times V$ matrix $\mathbf{B}$, where $\mathbf{B}^T = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]$. Each observation i is characterized by a latent vector $\boldsymbol{w}_i \in \Delta^{K-1}$. The elements $w_{i,k}$ of $\mathbf{w}_i$ represent the weight attached to $\boldsymbol{\beta}_k$ in generating $\mathbf{x}_i$. Hence, the count probabilities for observation i have the factor structure $\mathbf{p}_i = \sum_{k=1}^K \boldsymbol{\beta}_k w_{i,k} = \mathbf{B}^T \boldsymbol{w}_i$.[6] Combining these elements, the distribution of $\mathbf{x}_i$ is given by

$$\mathbf{x}_i | (C_i, \boldsymbol{w}_i) \sim \text{Multinomial}(C_i, \mathbf{B}^T \boldsymbol{w}_i), \tag{5}$$

where $C_i = \sum_{v=1}^V x_{i,v}$ is the count of all features in observation i. Finally, $\boldsymbol{\theta}_i$ is a sub-vector of $\boldsymbol{w}_i$ collecting the topic weights for inclusion in the regression.

Example: Monetary Policy Speeches. Suppose each unstructured observation is a monetary policy speech. One distribution $\boldsymbol{\beta}_k$ might place high weight on words like 'inflation', 'prices', and 'cpi', so $\boldsymbol{\beta}_k$ would have an interpretation as price rises. The corresponding $w_{i,k}$ represents how much speech i discusses price rises. One research question might ask how attention paid to price rises, along with economic conditions captured by numeric data, affects policy actions. This could be modeled by the ceofficient $\boldsymbol{\gamma}$ in (4), where Y_i is the policy action of speaker i, $\boldsymbol{\theta}_i$ selects the price rises topic weight from $\boldsymbol{w}_i$ and discards irrelevant topics, and $\mathbf{q}_i$ measures quantitative information like market forecasts for growth and inflation at the time the speech was made.

In the two-step strategy, one first constructs estimates $\hat{\mathbf{B}}$ of $\mathbf{B}$ and $(\hat{\boldsymbol{w}}_i)_{i=1}^n$ of $(\boldsymbol{w}_i)_{i=1}^n$ using LDA (Blei et al. 2003) or more recent methods (e.g., Bing et al. (2020), Wu et al. (2023), Ke and Wang (2022)). In the second step, Y_i is regressed on the sub-vector $\hat{\boldsymbol{\theta}}_i$ of $\hat{\boldsymbol{w}}_i$ and a vector of controls $\mathbf{q}_i$. Here the source of measurement error is sampling error in $(\hat{\boldsymbol{\theta}}_i)_{i=1}^n$. Each $\hat{\boldsymbol{\theta}}_i$ is estimated with a variance proportional to C_i^{-1}, so $\mathbb{E}[C_i^{-1}]$ controls

[6] This model nests as a special case a *pure multinomial* model where $K = V$, $\mathbf{B} = \mathbf{I}$, and $\boldsymbol{w}_i = \mathbf{p}_i$.

the overall rate of measurement error. The interplay between $\mathbb{E}[C_i^{-1}]$ and the number of observations n plays a key role in our theory.

3.1.3 Application 3: AI/ML-Generated Indices.

A third use case is the formation of indices through classification and aggregation. For example, Baker et al. (2016) construct EPU indices by classifying news articles as to pertaining to policy uncertainty then aggregating over time to form monthly or quarterly indices. Let C_i denote the sample of articles to be classified in month i, and X_i the number classified as pertaining to uncertainty. The quantity $\hat{\theta}_i = X_i/C_i$ provides a natural measure of the true latent uncertainty $\theta_i \in [0,1]$, where $\theta_i = 0$ represents no uncertainty and $\theta_i = 1$ represents maximal uncertainty. Here there is measurement error from both misclassification and sampling uncertainty. To account for these, we can frame this example as a topic model. Suppose the misclassification rates are constant across observations. Then $\mathbf{x}_i = (X_i, C_i - X_i)^T$ follows the distribution in (5), with

$$\mathbf{B}^T = \begin{bmatrix} \beta_1 & \beta_0 \\ (1-\beta_1) & (1-\beta_0) \end{bmatrix}, \qquad \boldsymbol{w}_i = \begin{bmatrix} \theta_i \\ 1-\theta_i \end{bmatrix},$$

where β_1 is the probability that an article pertaining to uncertainty is classified correctly, and β_0 is the probability that an article not pertaining to uncertainty is misclassified. The stylized example from Section 2 is a special case where $\beta_1 = 1$ and $\beta_0 = 0$, implying no misclassification.

3.2 Theory for the Two-Step Strategy

We first introduce some notation. Let

$$\boldsymbol{\psi} = \begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\alpha} \end{bmatrix}, \qquad \boldsymbol{\xi}_i = \begin{bmatrix} \boldsymbol{\theta}_i \\ \mathbf{q}_i \end{bmatrix}, \qquad \hat{\boldsymbol{\xi}}_i = \begin{bmatrix} \hat{\boldsymbol{\theta}}_i \\ \mathbf{q}_i \end{bmatrix}.$$

The OLS estimator of $\boldsymbol{\psi}$ in the two-step strategy is given by

$$\hat{\boldsymbol{\psi}} = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i Y_i \right). \tag{6}$$

The OLS estimators $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are the upper and lower blocks of $\hat{\boldsymbol{\psi}}$.

3.2.1 Fixed Population

We first consider the large-sample properties of $\hat{\boldsymbol{\psi}}$ as the number of observations grows ($n \to \infty$), while the distribution of $(Y_i, \boldsymbol{\xi}_i, \hat{\boldsymbol{\xi}}_i)_{i=1}^n$ remains fixed. This asymptotic frame-

work approximates empirical settings with relatively small amounts of unstructured data per observation but a large number of observations. In these settings, measurement error will dominate sampling error asymptotically, leading to inconsistency.

We verify the following conditions for AI/ML-generated labels and topic models in Appendix B. In what follows, let $\mathbf{0}$ denote a conformable vector or matrix of zeros.

Assumption 1. *(i)* $\mathbb{E}\left[\|\boldsymbol{\xi}_i\|^2\right] < \infty$, *and* $\mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]$ *has full rank.*

(ii) $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T \to_p \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]$, $\frac{1}{n}\sum_{i=1}^n (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T \to_p \mathbf{V}$ *with* $\mathbf{V}$ *a finite non-random symmetric matrix,* $\frac{1}{n}\sum_{i=1}^n (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\boldsymbol{\theta}_i^T \to_p \mathbf{W}$ *with* $\mathbf{W}$ *a finite non-random matrix,* $\frac{1}{n}\sum_{i=1}^n (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\mathbf{q}_i^T \to_p \mathbf{0}$, *and* $\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\xi}}_i \varepsilon_i \to_p \mathbf{0}$ *as* $n \to \infty$.

Assumption 1(i) is standard. Assumption 1(ii) requires that $(\hat{\boldsymbol{\xi}}_i, \boldsymbol{\xi}_i, \varepsilon_i)$ satisfy some laws of large numbers. Only the last two conditions in Assumption 1(ii) are substantive. They ensure that the measurement errors $\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i$ are uncorrelated with $\mathbf{q}_i$ and the regression errors ε_i are uncorrelated with $\hat{\boldsymbol{\xi}}_i$ asymptotically. These conditions can be relaxed, but doing so would complicate the bias expressions without altering our main point: the two-step strategy can lead to significant biases. Let $\boldsymbol{\Delta} = \mathbf{V} + \mathbf{W} + \mathbf{W}^T$.

Theorem 1. *Suppose that Assumption 1 holds. Then*

$$\hat{\boldsymbol{\psi}} \to_p \left(\mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right] + \begin{bmatrix} \boldsymbol{\Delta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right)^{-1} \left(\mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right] + \begin{bmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right) \boldsymbol{\psi}, \tag{7}$$

as $n \to \infty$, *provided the inverse exists. In particular, if* $\boldsymbol{\Delta}$ *and* $\mathbf{W}$ *are small,*

$$\text{plim}(\hat{\boldsymbol{\psi}}) = \boldsymbol{\psi} - \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]^{-1} \begin{bmatrix} (\mathbf{V} + \mathbf{W}^T)\boldsymbol{\gamma} \\ \mathbf{0} \end{bmatrix} + O\left(\|\boldsymbol{\Delta}\| \max\{\|\boldsymbol{\Delta}\|, \|\mathbf{W}\|\}\right). \tag{8}$$

Theorem 1 shows that $\hat{\boldsymbol{\psi}}$ is inconsistent due to measurement error in $\hat{\boldsymbol{\theta}}_i$. This is relevant for researchers using control variables derived from unstructured data, as OLS estimators of $\boldsymbol{\alpha}$ from regression of Y_i on $\mathbf{q}_i$ with controls $\hat{\boldsymbol{\theta}}_i$ are inconsistent when $\boldsymbol{\theta}_i$ and $\mathbf{q}_i$ are correlated. More constructively, Theorem 1 shows bias is proportional to the precision of $\boldsymbol{\theta}_i$. The matrix $\mathbf{V}$ represents the variance of measurement error, while $\mathbf{W}$ represents the covariance of measurement error and $\boldsymbol{\theta}_i$. In many cases, measurement error is "classical" ($\mathbf{V} > \mathbf{0}$, $\mathbf{W} = \mathbf{0}$), but in "non-classical" settings, such as latent binary labels (Aigner 1973), $\mathbf{W} \neq \mathbf{0}$. Expressions for $\mathbf{V}$ and $\mathbf{W}$ in the context of AI/ML-generated labels and topic models are derived in Appendix B. To summarize:

Application 1: AI/ML-Generated Labels. Let $\pi_i = \Pr(\hat{\theta}_i = 1|\mathbf{x}_i, \mathbf{q}_i)$ denote the classifier's probability of assigning label 1 given $(\mathbf{x}_i, \mathbf{q}_i)$ and let p_i denote the true probability $\Pr(\theta_i = 1|\mathbf{x}_i, \mathbf{q}_i)$. For Bayes classifiers π_i takes values in $\{0, 1\}$ whereas for randomized

classifiers it takes values in $[0, 1]$. The misclassification probability $M_i = p_i + \pi_i - 2p_i\pi_i$ is the probability that $\hat{\theta}_i \neq \theta_i$. It is the sum of the false-positive probability $FP_i = \pi_i(1-p_i)$ and false-negative probability $FN_i = p_i(1-\pi_i)$. Appendix B shows

$$\mathbf{V} = \mathbb{E}\left[M_i\right], \quad \mathbf{W} = -\mathbb{E}\left[FN_i\right].$$

Thus, $\mathbf{V}$ is the misclassification rate, which is the sum of the false-positive rate $\mathbb{E}[FP_i]$ and the false-negative rate $\mathbb{E}[FN_i]$. If both of these terms are small, Theorem 1 shows that the bias in $\boldsymbol{\gamma}$ is proportional to $\mathbf{V} + \mathbf{W}^T = \mathbb{E}[FP_i]$, the false-positive rate.

Application 2: Topic Models. Here we have

$$\mathbf{V} = \mathbb{E}\left[\frac{1}{C_i}\right]\left(\boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\,\mathrm{diag}(\mathbf{B}^T\mathbb{E}[\boldsymbol{w}_i])\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\boldsymbol{S}^T - \mathbb{E}\left[\boldsymbol{\theta}_i\boldsymbol{\theta}_i^T\right]\right), \quad \mathbf{W} = \mathbf{0},$$

where $\boldsymbol{\theta}_i = \boldsymbol{S}\boldsymbol{w}_i$ for a known matrix $\boldsymbol{S}$ and $\mathrm{diag}(\mathbf{v})$ is a diagonal matrix with $\mathbf{v}$ down its diagonal. The bias in $\boldsymbol{\gamma}$ is proportional to the average *inverse* amount of unstructured data per observation, $\mathbb{E}[C_i^{-1}]$. As this is an inverse relationship, bias may be large if most observations have large C_i but some have small C_i.

3.2.2 Sequence of Populations

In the previous subsection, measurement error dominated sampling error asymptotically. But in finite samples, both measurement error and sampling error will play a role. To better reflect the finite-sample problem faced in practice, we consider an alternative asymptotic framework in which both measurement error and sampling error matter. Formally, we consider a sequence of populations indexed by the sample size n, where the precision of $\hat{\boldsymbol{\theta}}_i$ increases with n. One interpretation of this framework is that it captures the prevailing trend whereby increasingly large datasets are analyzed by increasingly accurate algorithms. In each population, we keep the distribution of $(Y_i, \boldsymbol{\theta}_i, \mathbf{q}_i)$ held fixed and let the conditional distribution of $\hat{\boldsymbol{\theta}}_i$ given $(Y_i, \boldsymbol{\theta}_i, \mathbf{q}_i)$ vary with n, so that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T \to_p \kappa\,\boldsymbol{\Omega} \tag{9}$$

as $n \to \infty$, where $\boldsymbol{\Omega}$ is a finite non-random matrix and κ is a non-negative constant. Compared with expression (8), in (9) we allow $\mathbf{V}+\mathbf{W}^T$ to shrink with n, so that both bias and variance matter asymptotically. The constant κ represents the *relative* magnitudes of measurement error and sampling error. In Appendix B we show that (9) holds for AI/ML-generated labels and topic models and derive expressions for κ and $\boldsymbol{\Omega}$, illustrating how κ links measurement error in $\hat{\boldsymbol{\theta}}_i$ to the sample size n.

Application 1: AI/ML-Generated Labels. So that both misclassification and sampling uncertainty matter asymptotically, we let the false-positive rate shrink with n so that

$$\sqrt{n} \times \mathbb{E}[FP_i] \to \kappa.$$

Smaller values of κ correspond to settings with a smaller false-positive rate, and hence relatively less measurement error in the generated labels.

Application 2: Topic Models. To mimic common empirical settings where there is a large amount of unstructured data per observation and the number of observations is also large, let

$$\sqrt{n} \times \mathbb{E}\left[\frac{1}{C_i}\right] \to \kappa.$$

Smaller values of κ correspond to settings where there is a relatively more unstructured data per observation, and hence relatively less measurement error. As noted in the Introduction, empirical counterparts to κ range between 1 and 20 for a few widely-used unstructured data sets. Simulations reported in Section 5 show that two-step inference can suffer significant biases even for relatively small values of κ in this range.

In what follows, notions of convergence in probability and distribution should be understood as holding along this sequence of populations satisfying (9). We focus on the case where the data are independent and identically distributed or the error terms are martingale differences to simplify exposition, though the results can easily be extended to general types of dependence. Let $\hat{\varepsilon}_i = Y_i - \hat{\boldsymbol{\psi}}^T \hat{\boldsymbol{\xi}}_i$.

Assumption 2. *(i)* $\mathbb{E}\left[\|\boldsymbol{\xi}_i\|^2\right] < \infty$, $\mathbb{E}\left[\|\varepsilon_i \boldsymbol{\xi}_i\|^2\right] < \infty$, *and* $\mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]$ *has full rank.*

(ii) $\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T \to_p \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]$, $\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\boldsymbol{\theta}}_i(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T \to_p \kappa \boldsymbol{\Omega} \geq \mathbf{0}$, *with* $\boldsymbol{\Omega}$ *a finite non-random matrix and* κ *a non-negative constant, and* $\frac{1}{\sqrt{n}}\sum_{i=1}^n (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\mathbf{q}_i^T \to_p \mathbf{0}$ *as* $n \to \infty$.

(iii) $\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\boldsymbol{\xi}}_i \varepsilon_i \to_d N(\mathbf{0}, \mathbb{E}[\varepsilon_i^2 \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T])$ *and* $\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2 \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T \to_p \mathbb{E}[\varepsilon_i^2 \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T]$ *as* $n \to \infty$.

Assumption 2(i) is standard. Assumption 2(ii) strengthens Assumption 1(ii) to require convergence at a $\sqrt{n}$-rate. The first part of Assumption 2(iii) imposes a standard CLT condition while the second part is only used to establish consistency of standard errors. These conditions are verified for the running examples of AI/ML-generated labels and topic models in Appendix B.

Our second main result shows that $\hat{\boldsymbol{\psi}}$ is consistent, derives its asymptotic distribution, and establishes consistency of standard errors.

Theorem 2. *Suppose that Assumption 2 holds. Then*

$$\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \to_d N\left(-\kappa\, \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]^{-1} \begin{bmatrix} \boldsymbol{\Omega}\boldsymbol{\gamma} \\ \mathbf{0} \end{bmatrix}, \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]^{-1} \mathbb{E}\left[\varepsilon_i^2 \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right] \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]^{-1}\right), \quad (10)$$

and

$$\left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2 \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T\right) \left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T\right)^{-1} \to_p \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]^{-1} \mathbb{E}\left[\varepsilon_i^2 \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right] \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]^{-1}, \quad (11)$$

as $n \to \infty$.

Theorem 2 shows that two-step inference is *valid* when $\kappa = 0$. In this case, measurement error is of smaller order than sampling error asymptotically and so we have

$$\sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \to_d N\left(\mathbf{0}, \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]^{-1} \mathbb{E}\left[\varepsilon_i^2 \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right] \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]^{-1}\right).$$

Here $\hat{\boldsymbol{\psi}}$ has the same asymptotic distribution as the (infeasible) OLS estimator obtained by regressing Y_i on the true latent $\boldsymbol{\theta}_i$ and standard errors computed using $\hat{\boldsymbol{\theta}}_i$ are consistent.

Two-step inference is *invalid* when $\kappa > 0$, as both measurement error and sampling error are non-negligible. While OLS standard errors are consistent and the asymptotic distribution of $\hat{\boldsymbol{\psi}}$ is Normal with the same variance as the $\kappa = 0$ case, its center is shifted away from the origin due to measurement error bias. As a result, confidence intervals based on the usual two-step strategy have the correct width but incorrect centering, leading to coverage rates below nominal coverage.[7] The bias—and thus the degree of under-coverage—increases in κ. Simulations reported in Section 5 show that coverage distortions can be severe even for small values of κ. This critique applies to inference on $\boldsymbol{\alpha}$ as well as $\boldsymbol{\gamma}$, and is therefore relevant for researchers using unstructured data to create controls variables.

Remark 1. These implications contrast with a generated regressors problem, where the asymptotic variance is inflated but there is no location shift. In the classical generated regressor problem (Pagan 1984), the $\hat{\boldsymbol{\theta}}_i$ depend on a common parameter that is estimated in the first stage. This across-observation dependence causes the term in display (9) to converge to a random variable rather than a constant, leading to the variance inflation.

Remark 2. The problem we study is related conceptually to factor-augmented regressions. In its simplest form, latent factors $\mathbf{F}_t$ are imputed from a vector of N predictor variables $\mathbf{x}_t$ using PCA, then the estimated factors $\hat{\mathbf{F}}_t$ are used as covariates in a regression model. Bai and Ng (2006) show that using the estimated factors $\hat{\mathbf{F}}_t$ as covariates

[7] We omit discussion of the case $\kappa = +\infty$ where measurement error dominates sampling error. In that case, the coverage rates of standard OLS confidence intervals approach zero as n becomes large.

leads to valid inference provided $\sqrt{T}/N \to 0$, where T is the time-series dimension and N is the cross-sectional dimension. Their T is analogous to our n, and, within the context of topic models, their $1/N$ is analogous to our $\mathbb{E}[C_i^{-1}]$. Thus, their condition $\sqrt{T}/N \to 0$ is analogous to $\kappa = 0$. Gonçalves and Perron (2014) show that if $\sqrt{T}/N$ converges to a constant, which is analogous to $\kappa > 0$, then there is a bias that shifts the location of the asymptotic distribution. At an abstract level, Theorem 2 can be seen as generalizing this finding to a broad class of scenarios. Our applications to AI/ML-generated models and topic models are novel and do not follow from these existing works.

Remark 3. As noted in the Introduction, our asymptotic framework is related to an econometrics literature (e.g., Chesher (1991)) on "small" classical measurement error, where the variance of measurement error shrinks to zero at rate $n^{-1/2}$. Recently, Evdokimov and Zeleneev (2023) show how to bias-correct GMM estimators in this context, using instrumental variables to identify the measurement error variance. Their approach imposes no structure on the source of measurement error. Our setting is fundamentally different: measurement error arises due to first-stage estimation of $\boldsymbol{\theta}_i$, allowing us to characterize bias without an instrument and perform analytical bias correction. We also allow non-classical measurement error, which is important for the case of imputed labels.

4 How To Do Valid Inference

Having shown the bias inherent to the popular two-step approach, we now propose two methods for performing valid inference on $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$. The first constructs bias corrected estimators and confidence intervals. The second is a one-step strategy that jointly estimates the upstream and downstream regression models.

Both approaches have strengths and weaknesses. The bias correction is simple to implement and scalable. On the other hand, the explicit formulas for the bias we develop are for particular settings. While these encompass leading applications, they are not comprehensive. On the other hand, the one-step strategy is more flexible and can handle cases where a bias formula may be complex to develop. But it is more computationally demanding.

4.1 Bias Correction

Theorem 2 shows that the asymptotic bias of the two-step estimator $\hat{\boldsymbol{\psi}}$ takes the form

$$\boldsymbol{b} = -\kappa\, \mathbb{E}\Big[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\Big]^{-1} \begin{bmatrix} \boldsymbol{\Omega\gamma} \\ \mathbf{0} \end{bmatrix}.$$

We use this formula to construct bias-corrected estimators and confidence intervals (CIs) for $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$. Suppose we have estimators $\hat{\kappa}$ of κ and $\hat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$. We discuss how to construct valid estimators within the context of our running examples below. We may estimate the bias using

$$\hat{\boldsymbol{b}} = -\hat{\kappa}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1}\begin{bmatrix}\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\gamma}}\\ \mathbf{0}\end{bmatrix},$$

where $\hat{\boldsymbol{\gamma}}$ is the two-step estimator of $\boldsymbol{\gamma}$ (the upper block of $\hat{\boldsymbol{\psi}}$ in (6) corresponding to $\boldsymbol{\gamma}$). Let ψ_j denote the jth component of $\boldsymbol{\psi}$, $\hat{\psi}_j$ denote the jth component of $\hat{\boldsymbol{\psi}}$, and $\hat{b}_j$ denote the jth component of $\hat{\boldsymbol{b}}$. The bias-corrected estimator $\hat{\psi}_j^{bc}$ of ψ_j is

$$\hat{\psi}_j^{bc} = \hat{\psi}_j - \frac{\hat{b}_j}{\sqrt{n}}.$$

We may construct a valid $100(1-\alpha)\%$ CI for ψ_j by centering at the bias-corrected estimator:

$$\text{CI}(\psi_j) = \left[\hat{\psi}_j^{bc} - z_{1-\alpha/2}\frac{\hat{\sigma}_j}{\sqrt{n}}, \hat{\psi}_j^{bc} + z_{1-\alpha/2}\frac{\hat{\sigma}_j}{\sqrt{n}}\right],$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the normal distribution (e.g., 1.96 for a 95% CI), and $\hat{\sigma}_j$ denotes the square root of the jth diagonal entry of the covariance matrix estimator on the left-hand side of (11). The following result shows that $\text{CI}(\psi_j)$ is valid:

Theorem 3. *Suppose that Assumption 2 holds, that $\mathbb{E}[\varepsilon_i^2\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T]$ has full rank, and that $\hat{\kappa} \to_p \kappa$ and $\hat{\boldsymbol{\Omega}} \to_p \boldsymbol{\Omega}$. Then $\Pr(\psi_j \in \text{CI}(\psi_j)) \to 1-\alpha$ as $n \to \infty$.*

We now show how to bias-correct in the context of the two running examples. In both cases we propose consistent estimators $\hat{\kappa}$ and $\hat{\boldsymbol{\Omega}}$ of κ and $\boldsymbol{\Omega}$, from which it follows by Theorem 3 that bias-corrected CIs have correct coverage. We also suggest reporting $\hat{\kappa}$ as an easy-to-compute and useful diagnostic for understanding the relative importance of measurement error and sampling error.

Application 1: AI/ML-Generated Labels. Recall that $\kappa = \lim_{n\to\infty}\sqrt{n}\,\mathbb{E}[FP_i]$ and $\boldsymbol{\Omega} = 1$. Here κ can be estimated from a small validation sample in which both $\hat{\theta}_i$ and θ_i are observed (e.g., by human-labelling). We allow the size m of the validation data set to be of smaller order of magnitude than n, so that $m/n \to 0$, asymptotically. This is important for accommodating modern applications where ML/AI methods are deployed at scale to impute labels on massive data sets, but where correctly labeling data can be costly. For instance, Boxell and Conway (2022) impute binary labels representing political slant for a corpus of millions of newspaper articles using a validation sample size in the tens of thousands. Other approaches recently advocated in the literature for correcting measurement error in regressors generated using AI/ML (Fong and Tyler 2021,

Allon et al. 2023, Egami et al. 2023, Zhang et al. 2023) are only valid when m and n are of similar size (i.e., when $m/n \to c > 0$), making them ill-suited and potentially unreliable in modern applications where AI/ML methods are deployed at scale. Simulations reported in Section 5.1.1 support this claim.

We can estimate κ using

$$\hat{\kappa} = \sqrt{n}\widehat{FPR}, \qquad \widehat{FPR} = \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i(1-\theta_i).$$

In practice, the empirical false-positive rate $\widehat{FPR}$ can simply be read off a confusion table in a validation exercise. The following result shows that $\hat{\kappa}$ is consistent, even when the overall sample size n is of larger order of magnitude than m.

Lemma 1. *Suppose that* $\sqrt{n}\,\mathbb{E}[FP_i] \to \kappa > 0$ *and* $n/m^2 \to 0$. *Then* $\hat{\kappa} \to_p \kappa$.

Consistency of $\hat{\kappa}$ suffices for asymptotic validity of the bias-corrected CIs. However, the estimation of the false-positive rate from a small validation sample can introduce additional sampling variability that, while asymptotically negligible, can be important to account for in finite samples. To this end, we introduce the following finite-sample correction. Let $\hat{\mathbf{V}}$ denote the Eicker–Huber–White covariance matrix estimator in the left-hand side of display (11). Also let

$$\hat{\mathbf{\Gamma}} = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T\right)^{-1} \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

The adjusted covariance matrix estimator $\tilde{\mathbf{V}}$ is given by

$$\begin{aligned}\tilde{\mathbf{V}} = {} & (\mathbf{I} + \widehat{FPR}\,\hat{\mathbf{\Gamma}})\hat{\mathbf{V}}(\mathbf{I} + \widehat{FPR}\,\hat{\mathbf{\Gamma}})' + \frac{1}{m}\widehat{FPR}(1-\widehat{FPR})\hat{\mathbf{\Gamma}}\hat{\mathbf{V}}\hat{\mathbf{\Gamma}}' \\ & + \frac{n}{m}\widehat{FPR}(1-\widehat{FPR})\hat{\mathbf{\Gamma}}\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}'\hat{\mathbf{\Gamma}}'.\end{aligned}$$

The motivation for this adjustment follows from the law of total variance. Under the conditions of Theorem 2 and Lemma 1, these adjustments are asymptotically negligible and the estimator $\tilde{\mathbf{V}}$ is consistent for the asymptotic variance derived in Theorem 2.

In practice, we recommend constructing CIs with standard errors computed from $\tilde{\mathbf{V}}$. For instance, to construct a 95% CI for ψ_j, we recommend using

$$\mathrm{CI}(\psi_j) = \left[\hat{\psi}_j^{bc} - 1.96\frac{\tilde{\sigma}_j}{\sqrt{n}}, \hat{\psi}_j^{bc} + 1.96\frac{\tilde{\sigma}_j}{\sqrt{n}}\right],$$

where $\tilde{\sigma}_j$ is the square root of the jth diagonal entry of $\tilde{\mathbf{V}}$.

Application 2: Topic Models. Recall that $\kappa = \lim_{n\to\infty} \sqrt{n}\,\mathbb{E}[C_i^{-1}]$. Theorem 7 in Appendix B shows that $\boldsymbol{\Omega} = \boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\,\mathrm{diag}(\mathbf{B}^T\mathbb{E}[\boldsymbol{w}_i])\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\boldsymbol{S}^T - \mathbb{E}\left[\boldsymbol{\theta}_i\boldsymbol{\theta}_i^T\right]$. Let $\hat{\kappa} = \frac{1}{\sqrt{n}}\sum_{i=1}^n C_i^{-1}$ and

$$\hat{\boldsymbol{\Omega}} = \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\,\mathrm{diag}(\hat{\mathbf{B}}^T\bar{\boldsymbol{w}}_n)\hat{\mathbf{B}}^T(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\boldsymbol{S}^T - \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\theta}}_i\hat{\boldsymbol{\theta}}_i^T,$$

where $\bar{\boldsymbol{w}}_n = \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{w}}_i$. No new computations are required, as each constituent term of $\hat{\boldsymbol{\Omega}}$ is computed when implementing the two-step strategy. The following result shows that these estimators are consistent.

Lemma 2. *Suppose that Assumption 6 in Appendix B holds and that* $\bar{\boldsymbol{w}}_n \to_p \mathbb{E}[\boldsymbol{w}_i]$. *Then* $\hat{\kappa} \to_p \kappa$ *and* $\hat{\boldsymbol{\Omega}} \to_p \boldsymbol{\Omega}$.

4.2 One-Step Strategy

Our second approach for correcting the bias from the two-step strategy is to estimate a joint likelihood over the model for $\boldsymbol{\theta}_i$ and the regression model in (4). The strategy can be deployed generically, but we will discuss it in the concrete setting of the topic model.

The starting point is to note that the topic model (5) provides a likelihood for $\mathbf{x}_i$ conditional on C_i and $\boldsymbol{\theta}_i$. We next build a joint likelihood for $(\mathbf{x}_i, Y_i, \boldsymbol{\theta}_i)$ by specifying a parametric distribution for the regression errors in (4). Finally, in the spirit of correlated random effects estimation in panel data models, we also specify a distribution for the topic shares $\boldsymbol{\theta}_i$ conditional on covariates $\mathbf{g}_i$. The covariates $\mathbf{g}_i$ (J in total) may or may not be the same as $\mathbf{q}_i$. In practice, one must specify particular distributions to complete the specification of the likelihood, but for now we keep the discussion general to highlight the broad applicability of the approach. In Section 5.2 we present a specific model—the Supervised Topic Model with Covariates—that we use for the simulations and empirics.

Together, these components combine to give a likelihood $l(\mathbf{x}_i, Y_i, \boldsymbol{\theta}_i | C_i, \mathbf{g}_i, \mathbf{q}_i)$ for $\mathbf{x}_i$, Y_i, and $\boldsymbol{\theta}_i$ conditional on C_i and covariates $\mathbf{g}_i$ and $\mathbf{q}_i$. As $\boldsymbol{\theta}_i$ is latent, we can integrate it out to obtain a likelihood $l(\mathbf{x}_i, Y_i | C_i, \mathbf{g}_i, \mathbf{q}_i)$ depending only on observable variables, which can then be used for maximum likelihood estimation of model parameters $\boldsymbol{\delta}$, consisting of $\mathbf{B}$, $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$, and any other parameters in the regression error and topic share distributions. However, there are two challenges. First, the integration has no closed-form solution and so must be performed numerically. Moreover, this numerical integration is high-dimensional and must be done observation-by-observation. As such, standard likelihood-based estimation is not computationally feasible. In the remainder of this section, we discuss how we overcome this computational challenge.

4.2.1 Inference Approach for the One-Step Strategy

The inference approach is frequentist but uses Bayesian computation. Integrating $\boldsymbol{\theta}_i$ out of the likelihood is performed implicitly as part of the sampling procedure. In this approach, we introduce a prior for the model parameters $\boldsymbol{\delta}$ and treat the latent $\boldsymbol{\theta}_i$ as "parameters" drawn from a distribution that potentially depends on covariates $\mathbf{g}_i$, as discussed above. We sample from the posterior distribution of $(\boldsymbol{\delta}, (\boldsymbol{\theta}_i)_{i=1}^n)$ conditional on the observed data $(\mathbf{x}_i, Y_i, C_i, \mathbf{g}_i, \mathbf{q}_i)_{i=1}^n$. The marginal draws for $\boldsymbol{\delta}$ represent draws from the posterior distribution for $\boldsymbol{\delta}$ based on the integrated likelihood $l(\mathbf{x}_i, Y_i | C_i, \mathbf{g}_i, \mathbf{q}_i)$.

It is important to emphasize that while our approach uses Bayesian computation, one does in fact perform valid *frequentist* inference on model parameters $\boldsymbol{\delta}$ using this method. The maximum likelihood estimator $\hat{\boldsymbol{\delta}}$ of $\boldsymbol{\delta}$ is asymptotically normal under standard regularity conditions (e.g., Theorem 5.41 of van der Vaart 1998). By the Bernstein–von Mises Theorem (see Theorem 10.1 of van der Vaart 1998 and discussion), the posterior mean $\bar{\boldsymbol{\delta}}$ of $\boldsymbol{\delta}$ is first-order asymptotically equivalent to the MLE $\hat{\boldsymbol{\delta}}$. Moreover, the posterior distribution of $\boldsymbol{\delta}$ is asymptotically normal with mean $\bar{\boldsymbol{\delta}}$ and variance (when appropriately scaled with n) equal to the asymptotic variance of the MLE. As such, Bayesian credible sets for $\boldsymbol{\delta}$—or any of its components such as $\boldsymbol{\gamma}$—are valid frequentist confidence sets with the desired asymptotic coverage. This approach is also *efficient* for inference on $\boldsymbol{\delta}$ and its components, as it is asymptotically equivalent to likelihood-based inference.

4.2.2 Hamiltonian Monte Carlo

Our problem is to sample from the posterior distribution $q(\boldsymbol{\zeta} | (\mathbf{x}_i, Y_i, C_i, \mathbf{g}_i, \mathbf{q}_i)_{i=1}^n)$ where $\boldsymbol{\zeta} = (\boldsymbol{\delta}, (\boldsymbol{\theta}_i)_{i=1}^n)$. To do so, we use Hamiltonian Monte Carlo (HMC), a modern Markov chain Monte Carlo (MCMC) algorithm that is particularly well-suited to high-dimensional models.[8] MCMC algorithms define a stochastic process, i.e., a Markov chain, whose ergodic distribution coincides with the posterior distribution one wishes to sample from. Samples from this Markov chain can be used to form estimates and confidence intervals. Efficient MCMC algorithms have low autocorrelation across samples which improves the accuracy of the resulting estimates.

A popular and simple MCMC method is the Metropolis-Hastings (MH) algorithm. Note the posterior is proportional to $q_n(\boldsymbol{\zeta}) := q(\boldsymbol{\zeta}, (\mathbf{x}_i, Y_i)_{i=1}^n | (C_i, \mathbf{g}_i, \mathbf{q}_i)_{i=1}^n)$, which is formed by multiplying the likelihood by the prior. The MH algorithm generates samples from the posterior in two steps: (i) propose a new state $\boldsymbol{\zeta}'$ from the current state $\boldsymbol{\zeta}$ using a pre-specified proposal distribution; then (ii) accept the new proposal with a probability that increases in the ratio $q_n(\boldsymbol{\zeta}')/q_n(\boldsymbol{\zeta})$. A challenge in practice is that the proposal

[8] More in-depth overviews of HMC are provided in Neal (2012), Hoffman and Gelman (2014), and Betancourt (2018). We are not aware of the application of HMC to topic models in the literature.

distribution must be chosen carefully to avoid slow convergence. Taking small steps in a random direction can have a high acceptance probability but also high autocorrelation across samples and slow convergence. Taking a large step in a random direction can drastically reduce $q_n(\boldsymbol{\zeta}')$ and hence the acceptance probability.

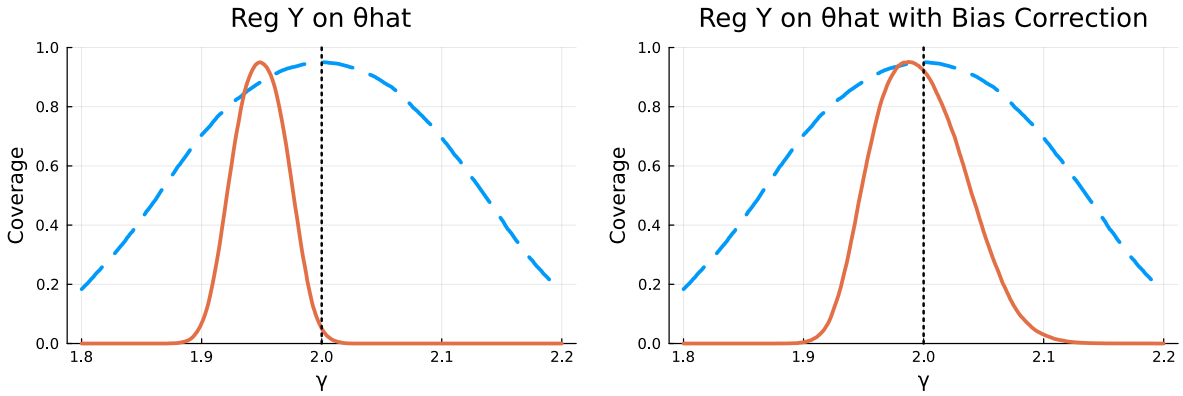
HMC uses the geometry of q_n to propose distant states that have high chance of acceptance. A new state $\boldsymbol{\zeta}'$ is proposed by following Hamiltonian dynamics for a certain number of steps, starting from the initial state $\boldsymbol{\zeta}$. This process is determined by the curvature of q_n, and so determining the path to follow requires evaluating the gradient of q_n with respect to the parameters $\boldsymbol{\zeta}$. The specific variant of HMC that we use is the No-U-Turn Sampler (Hoffman and Gelman 2014, NUTS). Intuitively, NUTS follows Hamiltonian dynamics for a random number of steps, and stops when the path starts to double back on itself. This is not only more efficient than following the dynamics for a fixed number of steps, but also avoids the need to specify the number of steps in advance.

4.2.3 Implementation with Probabilistic Programming

From an implementation perspective, an advantage of HMC is that it is amenable to probabilistic programming. This allows one to define a data generating process for a statistical model in computer code, after which sampling is performed "automatically" in the background by following a generic set of algorithmic procedures adapted to the given model. In practice, modern probabilistic programming libraries use automatic differentiation to compute the gradients of highly flexible families of densities. Furthermore, the density and gradient computations are typically parallelizable as they are additive with respect to the data points.[9] This facilitates the use of the same specialized hardware normally used for machine learning tasks.

NUTS is implemented in many probabilistic programming libraries, the most popular of which is Stan. For this paper, we instead use NumPyro (Phan et al. 2019), which utilizes a state-of-the-art automatic differentiation engine Jax (Bradbury et al. 2018) and allows users to easily deploy these computations to specialized hardware such as Graphical Processing Units (GPUs) and Tensor Processing Units (TPUs), resulting in a dramatic improvement in computation time. Furthermore, NumPyro is a Python library, not a standalone program, which means that it is easy to integrate with other libraries and benefits from the host of functionalities that Python provides. This said, our goal is not to advocate for any particular library, but to demonstrate that software and hardware have evolved to a point that allows Bayesian computation to be performed at scale without the need to manually derive sampling equations.

[9] More precisely, the logarithm of q_n is additive with respect to the data points, and the gradient of the logarithm of q_n is the sum of the gradients of the log-likelihood and the logarithm of the prior.



(a) Two-Step CIs **(b)** Bias-Corrected CIs

Note: Each mountain plot presents the share of simulations in which the value of γ on the x-axis is included in the 95% confidence interval. The dotted vertical lines show the true value of the parameter. Blue dashed lines are for regression of Y on the true θ using the human-labeled subsample.

Figure 1: Coverage of Two-Step and Bias-Corrected CIs

5 Empirical Results

In this section, we evaluate the performance of the two-step strategy against our proposed alternatives for valid inference. We observe a meaningful differences in parameter estimates and confidence intervals produced by the two-step method and the valid methods we propose. These findings are in line with the key predictions of our theory.

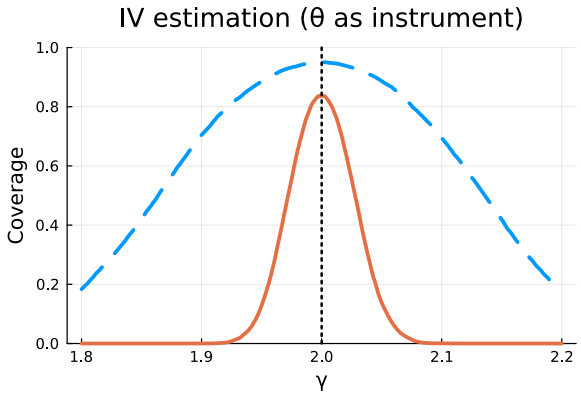
5.1 AI/ML-Generated Labels

5.1.1 Simulation Evidence

We consider a simulation environment with the following specification:

- Total number of observations is $n = 25,000$.
- Total number of human-labeled observations is $m = 1,000$.
- True label $\theta_i \in \{0, 1\}$ independently drawn with $\Pr(\theta_i = 1) = 0.54$.
- Observed label $\hat{\theta}_i$ is noisy measure of θ_i where $\Pr(\theta_i \neq \hat{\theta}_i) = 0.012$.
- True model is $Y_i = \alpha + \gamma\theta_i + \varepsilon_i$, where $\alpha = 0$, $\gamma = 2$, and $\varepsilon_i \sim t_{12}$.

We plot coverage of 95% CIs as a function of γ for the two-step strategy in Figure 1a and the bias-corrected CIs in Figure 1b. The blue dashed curves in Figures 1a and 1b



Note: Each mountain plot presents the share of simulations in which the value of γ on the x-axis is included in the 95% confidence interval. The dotted vertical lines show the true value of the parameter. Blue dashed lines are for regression of Y on the true θ using the human-labeled subsample.

Figure 2: Coverage of IV-Based CIs

present coverage of 95% CIs from regressing Y_i on θ_i for the validation sample of human-labeled data of size $m = 1,000$. The estimator has correct coverage at the true $\gamma = 2$ but is noisy due to the relatively small sample. Consequently, confidence intervals are unnecessarily wide and coverage remains high far from the truth.

The orange solid curve in Figure 1a reports coverage for two-step CIs obtained by regressing Y_i on $\hat{\theta}_i$ for the full sample of size $n = 25,000$. The two-step CIs have highest coverage away from the truth, reflecting the incorrect centering of two-step CIs due to measurement-error bias, as predicted by Theorem 2. The coverage at the true value of $\gamma = 2$ is poor—around 5% instead of 95%—even though the measurement error bias is relatively small in absolute terms.

Figure 1b reports coverage of bias-corrected CIs from Section 4.1. We construct these by centering at the bias-corrected estimator $\hat{\gamma}^{bc}$ of γ and using the finite-sample adjustment to standard errors for AI/ML generated labels described in Section 4.1. Bias correction works: bias-corrected CIs achieve highest coverage close to the true value of γ, illustrating that they are centered near the truth. Moreover, the coverage of bias-corrected CIs at the true value of $\gamma = 2$ is close to nominal coverage of 95% and competitive with the regression involving only human labels. But, since the full sample is used in estimation, the bias-corrected CIs are much narrower and their coverage for values of γ away from the truth is substantially lower.

Several papers (Fong and Tyler 2021, Allon et al. 2023, Egami et al. 2023, Zhang et al. 2023) have recently advocated IV strategies for addressing measurement-error bias for regression with AI/ML-generated variables. The basic idea is to estimate a first-stage regression of θ_i on $\hat{\theta}_i$ using the validation sample of size m, then regress Y_i on the

fitted values in the full sample of size n. The asymptotic environment that justifies this approach assumes that $m/n \to c$ for some constant $c > 0$, as in the classical literature on auxiliary data (see Chen et al. 2008, for example). While the approaches differ, these methods all produce CIs that contract at rate $n^{-1/2}$. In modern applications, however, the sample of human-labeled observations is usually small relative to the full sample. In these cases, the estimation uncertainty of the first-stage model, which is of order $m^{-1/2}$, dominates sampling error from the second-stage regression. As such, when m is small relative to n, IV-based CIs are misleadingly narrow and will under-cover.

Since m is small relative to n in this simulation, one would expect the IV approach to perform poorly. Figure 2 shows this is indeed the case, with low coverage at $\gamma = 2$. On the other hand, the bias-corrected CIs perform well even when m is smaller than n. This illustrates the robustness of our approach in settings where m is small relative to n, and the importance of having theoretical guarantees that allow $m/n \to 0$.

5.1.2 Empirical Application: Remote Work

Since the COVID-19 pandemic, the incidence of remote work has risen remarkably (Barrero et al. 2021, Aksoy et al. 2022). But much of the evidence on remote work comes from surveys which are limited in sample size. This makes tracking the evolution of remote work across narrow geographies and firms infeasible. Hansen et al. (2023) instead develops a dataset (available at `https://wfhmap.com/`) that measures remote work from a vast corpus of online job postings provided by Lightcast. Each job posting contains structured metadata that provides information on occupation, firm, location, job title, and so forth. It also contains a textual description of the job that provides many more details in unstructured format.

Hansen et al. (2023) uses the posting text to impute a binary label to the posting. The classification begins with a sample of 10,000 paragraphs from the dataset, each of which is labeled by three separate individuals recruited from Amazon Mechanical Turk. The authors use fine-tuned DistilBERT (Sanh et al. 2020) to predict the labels given text. To evaluate the model's performance, the model is trained on 17,850 human labels (corresponding to 5,950 texts) and used to generate a predicted label for 4,050 test-set postings. The ground truth labels for the test set come from the majority human label. The resulting confusion matrix is displayed in Table 1. The classifier is quite accurate, with small false positive and false negative rates.

Hansen et al. (2023) uses the classifier to impute a remote work indicator for each observation in the entire corpus. The resulting dataset can be used to answer numerous questions related to the causes and consequences of remote work. One question of growing interest is the degree of inequality in remote work arrangements, which the data is well suited to answer (Lambert et al. 2023). At the same time, the measurement error in the

Table 1: Confusion Matrix for Remote Work Classifier

		Predictions	
		Negative	Positive
Human Labels	Negative	2,878 (71.1%)	39 (1.0%)
	Positive	34 (0.8%)	1,099 (27.1%)

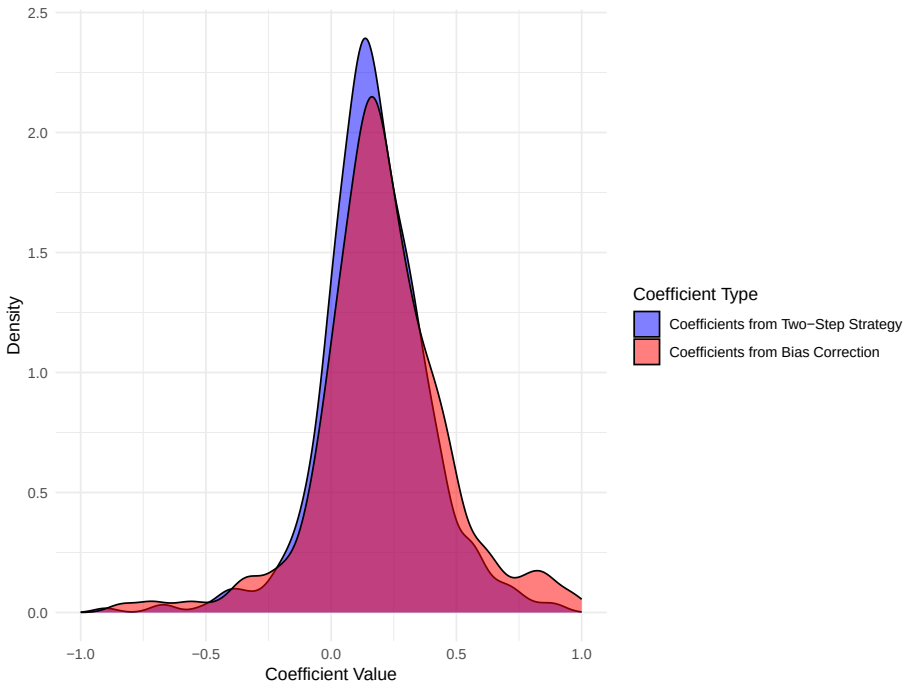


Figure 3: Distributions of Regression Coefficients of Wages on Remote Work Indicator by Occupation

indicator, although small, can still distort inference.

To provide an initial illustration of the consequences of two-step estimation for inference, we consider a sample of all US job postings in January 2022. The number of postings with a wage observation is 1,268,651. We first regress log posted wages on the remote indicator by SOC6 occupation and ignore measurement error. We then apply our bias correction formula to adjust the point estimates and confidence intervals. Figure 3 compares the two procedures. In general, there is a positive association between wages and remote work, suggesting that wage inequality is positively correlated with remote work inequality. But the quantitative impact is notably higher with the bias correction, in spite of the false positive rate being low: the average absolute value of the coefficient without correction is 0.226 and with correction is 0.337. The bias correction increases the coefficient on average in absolute terms by 47%. Bias correction also has dramatic effects on inference. Only 17.3% of the two-step coefficients fall in the bias-corrected 95%

confidence intervals.

5.2 Topic Models

To implement the one-step strategy for the topic model, we specify distributional assumptions to fully formulate a likelihood for empirical analysis. We assume the regression errors in (4) are normally distributed and that the distribution of $\boldsymbol{\theta}_i$ conditional on $\mathbf{g}_i$ is logistic normal. These assumptions are made for illustrative purposes and applied researchers may modify them as desired. The resulting model, which we call the *Supervised Topic Model with Covariates* (STMC), is formalized in Model 1.

$$\begin{aligned} \boldsymbol{\theta}_i &\sim \text{LogisticNormal}\left(\boldsymbol{\Phi}\mathbf{g}_i, \mathrm{I}_K\sigma_\theta^2\right) \\ \mathbf{x}_i &\sim \text{Multinomial}\left(C_i, \mathbf{B}^T\boldsymbol{\theta}_i\right) \end{aligned} \qquad \text{(Upstream Topic Model)}$$

$$Y_i \sim \text{Normal}\left(\boldsymbol{\gamma}^T\boldsymbol{\theta}_i + \boldsymbol{\alpha}^T\mathbf{q}_i, \sigma_Y^2\right) \qquad \text{(Downstream Regression Model)}$$

Model 1: Supervised Topic Model with Covariates

The matrix $\boldsymbol{\Phi}$ is a $K \times J$ matrix of regression coefficients. The kth row of $\boldsymbol{\Phi}$, denoted $\boldsymbol{\phi}_k$, captures how variation in covariates maps to variation in the prevalence of the kth topic across observations. Hence, a number of research questions can be addressed by performing inference on $\boldsymbol{\Phi}$. Model 1 also introduces scale parameters σ_θ and σ_Y. While we have modeled the error terms in the downstream regression and upstream logistic normal as homoskedastic to simplify presentation, this can easily be relaxed.

Example: Monetary Policy Speeches (Continued). To return to the example of Section 3, the downstream regression model could capture how policymakers' attention predicts policy actions controlling for economic conditions. But policymakers' attention can itself be a function of speaker characteristics such as demographic variables, or past experience of economic conditions (Malmendier et al. 2021). Such variables would enter $\mathbf{g}_i$ but arguably not directly affect policy decisions beyond their effect on attention; i.e., they would not enter $\mathbf{q}_i$.

To our knowledge, STMC is new in the literature. Roberts et al. (2014) presents a model in which a logistic normal distribution over $\boldsymbol{\theta}_i$ is parameterized by covariates but without a downstream regression. Blei and McAuliffe (2010) and Ahrens et al. (2021) present models in which linear combinations of topic shares explain a normally distributed response variable, but do not allow covariates to enter the distribution over $\boldsymbol{\theta}_i$. As such, we view STMC as of independent interest in the literature on topic modeling, although its

primary purpose is to provide an example in which dimensionality reduction and linear regression are part of the same joint model and one cares about doing valid inference on model parameters.

Following the literature on topic modeling, we specify the following standard prior distributions for model parameters:

$$\begin{aligned}
\boldsymbol{\beta}_k &\sim \text{Dirichlet}(\eta) \ \forall \ k \\
\phi_{j,k} &\sim \text{Normal}(0, \sigma_\phi^2) \ \forall \ j, k \\
\gamma_k &\sim \text{Normal}(0, \sigma_\gamma^2) \ \forall \ k \\
\alpha_m &\sim \text{Normal}(0, \sigma_\alpha^2) \ \forall \ m \\
\sigma_Y &\sim \text{Gamma}(s_0, s_1)
\end{aligned} \tag{Priors}$$

In total, the model has seven hyperparameters: the three σ^2 terms in (Priors) as well as σ_θ^2 in (Upstream Topic Model); the symmetric Dirichlet parameter η in (Priors); the two Gamma distribution parameters in (Priors).

Appendix D displays the NumPyro code needed to sample from the posterior distribution of STMC. The core code is only several dozen lines long, and individual elements can be quickly modified to specify alternative distributions, priors, or models without having to re-derive complex inference algorithms every time the model is adjusted.

In all both the simulation and application below, we use HMC with hyperparameters detailed in Appendix C. We choose $K = 2$, which implies that each observation's topic share vector can be written $\boldsymbol{\theta}_i = (\theta_i, 1-\theta_i)$. For the *one-step strategy*, we sample from the posterior distribution implied by the full structure of STMC. For the *two-step strategy*, we first sample from (Upstream Topic Model) and include only a constant in $\mathbf{g}_i$. We use the sampled values of θ_i to compute an estimate $\hat{\theta}_i$ of the posterior mean. We then estimate the following regression models using HMC:

$$\log\left(\frac{\hat{\theta}_i}{1-\hat{\theta}_i}\right) = \boldsymbol{\phi}^T \mathbf{g}_i + u_i, \tag{12}$$

$$Y_i = \gamma \hat{\theta}_i + \boldsymbol{\alpha}^T \mathbf{q}_i + \varepsilon_i, \tag{13}$$

where the error terms are normal. The prior distributions over the regression coefficients are the same in both strategies. This procedure is designed to emulate the typical approach in the empirical literature while ensuring that any observed differences between the two strategies are not driven by different inference methods or modeling choices.

Finally, our focus here is on inference rather than identification. Ke et al. (2021) highlight that the parameters of topic models are generally set- rather than point-identified. To restore point identification, a common assumption in the ML literature is the existence

of "anchor words" (Arora et al. 2012) which we adopt as explained below.[10]

5.2.1 Simulation Evidence

We simulate the data according to the data generating process described in Model 1.[11] We conduct three sets of 200 simulations of size n =10,000. Within each set, we let $C_i = C \in \{10, 25, 200\}$, which implies $\kappa \in \{10, 4, 0.5\}$, respectively. Further details are presented in Appendix C. We focus on the estimation of γ, the effect of the increase in θ_i on Y_i, and ϕ, the effect of a covariate g_i in (12). To see that the difference between the one-step and two-step strategies is due to mis-measurement of θ_i, we also report two-step estimates using the true latent θ_i as an input instead of $\hat{\theta}_i$. This approach is, of course, infeasible in practice.

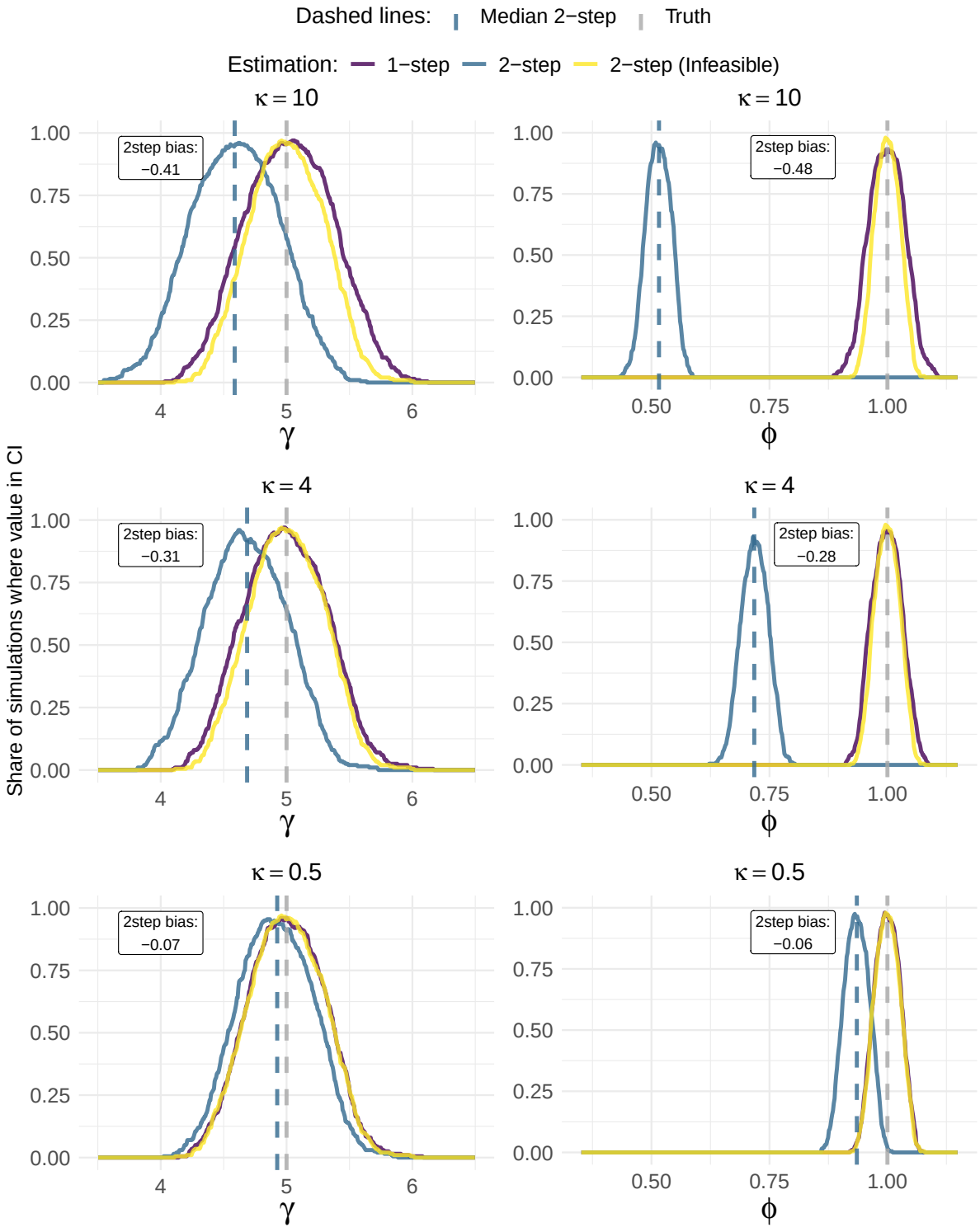
Figure 4 presents the results. Each panel shows the coverage rates of 95% confidence intervals for the different inference methods. The grey vertical dashed lines show the true value of the parameter. The blue vertical dashed line represents the median (across simulations) of mean posterior estimates for the two-step strategy. The two top panels show the results for the set of with $\kappa = 10$. As our theory predicts, the two-step strategy performs poorly in this case. The median (across simulations) estimate of γ is substantially biased towards zero. As predicted by theory, the width of two-step CIs is similar to the infeasible estimator that uses the true θ_i. Consequently, two-step CIs under-cover: Table 2 shows coverage for the true γ when $\kappa = 10$ is only 57.5%. Moving down the panels, we see that the bias decreases and coverage increases as the amount of unstructured data C increases or, equivalently, as κ decreases. On the other hand, the one-step strategy performs very well: the estimates appear unbiased, and the CIs have close to nominal coverage. The difference between the lengths of CIs using the one-step strategy and those using the infeasible estimator is small and decreasing in κ.Overall, the simulations confirm the three main insights from Theorem 2: (i) two-step estimates have a first-order bias that is driven by mis-measurement of θ_i; (ii) the bias is increasing in κ; and (iii) the width of two-step confidence intervals is not substantially affected by mis-measurement of θ_i.

5.2.2 Empirical Application: CEO Behavior

Bandiera et al. (2020) collects and analyzes data on CEO time use in a sample of manu-

[10] An alternative approach would be to dispense with the anchor words assumption, thereby allowing for the possibility of partial identification, and use an identification-robust method for constructing confidence sets based on the HMC draws as in Chen et al. (2018).

[11] We impose the anchor word assumptions in the simulation in the following way. We first draw $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ from symmetric $(V-50)$–dimensional Dirichlet priors. Then we insert 50 zeros into both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in such that there is no feature v where $\beta_{v,1} = \beta_{v,2} = 0$. Data is then simulated from these modified topic-feature distributions.



Note: Each mountain plot presents the share of simulations in which the value of γ and ϕ on the x-axis is included in the 95% confidence interval. The grey vertical dashed lines show the true value of the parameter. The blue vertical dashed line represents the median (across simulations) of mean posterior estimates from the two-step strategy. The bias reported is the difference between the truth and this median value.

Figure 4: Evolution of Bias in Regression Coefficients across κ Values

Table 2: Coverage Rates of Confidence Intervals

	(a) Coverage for γ			(b) Coverage for ϕ		
κ	2-Step	1-Step	Infeasible	2-Step	1-Step	Infeasible
10	0.575	0.955	0.955	0.000	0.920	0.975
4	0.635	0.965	0.955	0.000	0.955	0.975
1	0.850	0.975	0.955	0.000	0.945	0.975
0.5	0.910	0.960	0.955	0.025	0.965	0.975

Note: This table reports the coverage rates of 95% confidence intervals for γ and ϕ across different values of κ. The values are reported for the two-step and one-step strategies, as well as for the infeasible estimator that uses the true θ_i.

facturing firms. Their goal is to describe salient differences in executive time use, and to relate those differences to firm and CEO characteristics and firm outcomes.

The estimation sample consists of 916 CEOs, each of whom participated in a survey that recorded features of time use in each 15-minute interval of a given week, e.g. Monday 8am-8:15am, Monday 8:15am-8:30am, and so forth. The recorded categories are (1) the type of activity (meeting, public event, etc.); (2) duration of activity (15m, 30m, etc.); (3) whether the activity is planned or unplanned; (4) the number of participants in the activity; (5) the functions of the participants in the activity (HR, finance, suppliers, etc.). In total there are 654 unique combinations of these categories observed in the data. We let $x_{i,j}$ denote the number of times feature j appears in the time use diary of CEO i. The average value of C_i is 88.4, with a minimum of 2 and a maximum of 222. Bandiera et al. (2020) uses LDA with $K = 2$ to organize the time use data. The authors refer to the separate distributions over time use combinations $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ as *pure behaviors*. The share of CEO i's time devoted to pure behavior 1, θ_i, is referred to as the *CEO index*.

The authors use the following inference procedure. First, they estimate LDA on the time use data and set $\hat{\theta}_i$ to be the posterior mean of θ_i. They then use $\hat{\theta}_i$ as an input into productivity regressions where Y_i is the log of firm i sales, and $\mathbf{q}_i$ is a vector of firm observables. Further, they separately analyze which CEO and firm characteristics are associated with behaviors by regressing $\hat{\theta}_i$ on a vector of characteristics $\mathbf{g}_i$.

We re-examine these questions using the Supervised Topic Model with Covariates. To explain CEO behavior, in $\mathbf{g}_i$ we include log employment (a measure of firm size) and an indicator for whether the CEO has an MBA degree. To explain sales, in $\mathbf{q}_i$ we include log employment and fixed effects for year and country. As before, we use HMC for inference and the same priors for both strategies.[12] The priors used are the same as in the simulation exercise, except we set $\eta = 0.1$ as in Bandiera et al. (2020).

[12]We impose the anchor word assumption by zeroing out from $\boldsymbol{\beta}_1$ ($\boldsymbol{\beta}_2$) the activity that is relatively least likely in Pure Behavior 1 (2).

Table 3: Comparison of Pure Behaviors

Activity	1-step	2-step	Bandiera et al (2020)
Plant Visits	0.1	0.09	0.11
Suppliers	0.61	0.74	0.32
Production	0.38	0.33	0.46
Just Outsiders	0.74	1.21	0.58
Communication	1.44	1.23	1.49
Multi-Function	1.35	1.12	1.9
Insiders and Outsiders	1.8	1.83	1.9
C-suite	29.78	16.76	33.9

Note: This table reports the relative probability of certain activities in Pure Behavior 1 versus Pure Behavior 2. Values exceeding 1 indicate the activity is more likely to be performed under Pure Behavior 1. Values are reported in columns (1) and (2) are computed by first obtaining mean posterior probabilities of each activity in the given types. Column (3) presents values reported in Bandiera et al. (2020).

The key quantity that governs the relative importance of sampling error and measurement error is κ. In the context of the CEO behavior data, its empirical analog is 0.44, which is close to the lowest value of κ in the simulation exercise. This suggests that the two-step approach should perform relatively well. To further test our theory, we also estimate the model using a subsample of 10% of the activities for each CEO. This scenario could represent a researcher observing only half of a workday for each CEO, instead of a full five-day workweek. Such sampling increases the empirical analogue of κ to 4.26, which is near the middle value of κ in the simulation exercise.

Table 3 reports the relative probability of observing activities in Pure Behavior 1 versus Pure Behavior 2. Results obtained with the one- and two-step strategies are very similar, and are also similar to those reported in the original paper. The table suggests that interacting with C-Suite executives, spending time communicating, and holding multi-function meetings are much more likely under Pure Behavior 1. Conversely, spending time on plant visits and interacting solely with suppliers are more likely under Pure Behavior 2. Based on these observations, the original authors label the CEOs with high values of $\hat{\theta}_i$ as *leaders* and those with low values as *managers*.

Estimates reported in Table 4 are consistent with our theory and simulation results. In Panel (a), we show the estimates for the downstream regression coefficients, and in Panel (b), we show the estimates for the upstream coefficients. Columns (1) and (2) report the estimates obtained using one- and two-step strategies, respectively, for the full sample. Both strategies produce similar estimates and CIs for the coefficient of the CEO index in the downstream model, and suggest that a larger share of time spent on

Table 4: Regression Coefficient Estimates under Alternative Model Specifications

	Dependent variable: Log(sales)			
	(1) 2-Step	(2) 1-Step	(3) 2-Step	(4) 1-Step
CEO Index	0.4	0.402	0.211	0.439
	(0.219, 0.572)	(0.240, 0.603)	(-0.028, 0.449)	(0.153, 0.711)
Log Employment	1.212	1.198	1.239	1.199
	(1.159, 1.268)	(1.154, 1.248)	(1.186, 1.29)	(1.148, 1.26)
Controls	X	X	X	X
Activities' Sample	Full	Full	10%	10%

(a) Downstream Model: CEO Index and Firm Productivity

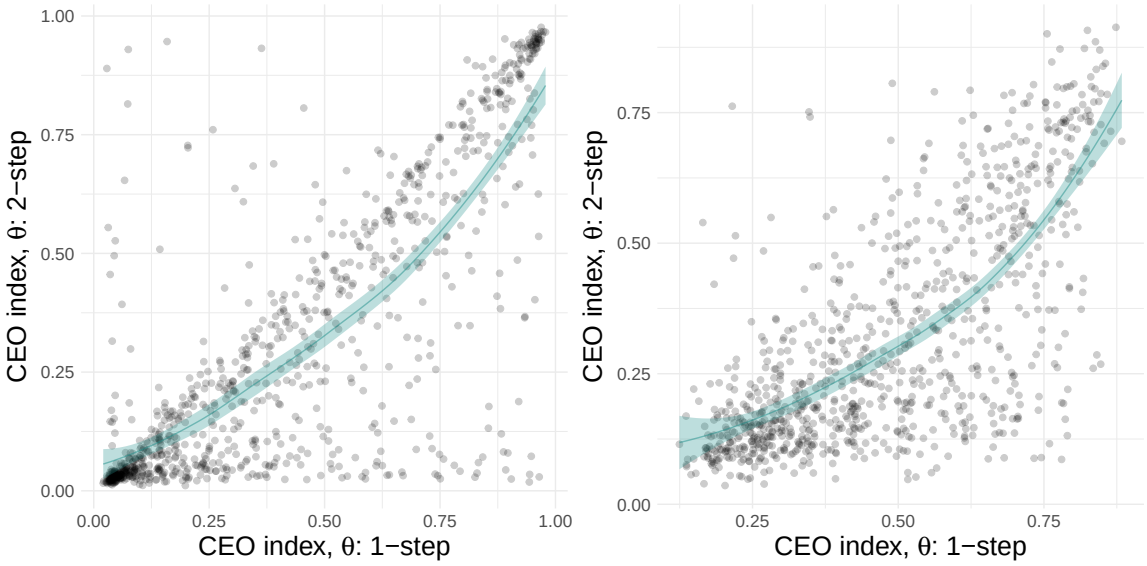
	Dependent variable: Un-normalized CEO index			
	(1) 2-Step	(2) 1-Step	(3) 2-Step	(4) 1-Step
MBA	0.307	0.606	0.118	0.323
	(0.176, 0.437)	(0.446, 0.743)	(-0.012, 0.249)	(0.107, 0.486)
Log Employment	0.356	0.492	0.154	0.443
	(0.306, 0.406)	(0.432, 0.548)	(0.104, 0.204)	(0.376, 0.507)
Controls	X	X	X	X
Activities' Sample	Full	Full	10%	10%

(b) Upstream Model: MBA and CEO Index

Note: 95% confidence intervals are shown in parentheses.

Pure Behavior 1 is associated with higher firm productivity. In the upstream model, we see larger differences between the two strategies as in the simulations. While having an MBA and managing larger firms are both associated with a higher CEO index, the point estimates differ substantially. There appears to be a downward bias in the two-step strategy: the coefficient on the MBA dummy is equal to 0.307 in the two-step strategy, compared to 0.606 in the one-step strategy. Moreover, there is no overlap in the CIs for these coefficients: the one-step CIs lie entirely to the right of the two-step CIs.

The differences between the strategies are substantially more pronounced when we instead use the 10% subsample of unstructured data. Under the one-step strategy, the empirical conclusions are largely the same as when using the full data. For example, the point estimate of γ changes from 0.402 to 0.439 and the confidence intervals, while wider, still indicate a positive relationship between CEO behavior and firm performance. This is not so with the two-step strategy: the point estimate of γ is now halved to 0.211, and the two-step CI includes 0. Likewise, in the upstream model, the estimate of the coefficient on the MBA indicator remains large and statistically significant in the one-step strategy, but is reduced by 62% and is no longer statistically significant in the two-step strategy. This is consistent with the theory and simulation results, which suggest that the two-step



(a) Activities' Sample: Full **(b)** Activities' Sample: 10%

Note: Each point represents the mean posterior estimate of a single CEO's index, $\hat{\theta}_i$. The blue line is the local polynomial fit (with confidence intervals) obtained with 'ggplots's' 'geom_smooth' with default parameters.

Figure 5: Scatterplots of Estimated CEO Indices $\hat{\theta}_i$

strategy should perform poorly with larger κ.

What explains the differences in estimates across strategies? To answer this question, we plot the estimated CEO indices in Figure 5. Panel (a) plots the estimated CEO indices obtained using the full sample, while Panel (b) plots the estimated CEO indices obtained using the 10% subsample. The blue line represents the local polynomial fit. Evidently, when the full sample is used, both strategies find a large number of CEOs with $\hat{\theta}_i$ close to 0 and 1, and a strong correlation between the two estimates. However, the correlation is much weaker for the 10% subsample, suggesting that there is a large scope for mis-measurement of θ_i.

6 Conclusion

The leading strategy for analyzing unstructured or high-dimensional data uses two steps. First, latent variables of economic interest are estimated using an AI-powered information retrieval algorithm or other ML method. Second, the AI- or ML-generated variables are plugged-in to downstream econometric models, and are treated as regular numeric "data" for the purposes of estimation and inference. This paper highlights, both theoretically and empirically, a previously unrecognized problem with this popular two-step strategy: measurement error introduced in the first step leads to biased estimates and invalid

inference for the downstream regression coefficients. The degree of bias, and therefore the degree to which it distorts inference, depends on the relative importance of measurement error and sampling error, but it can be material in applications. To guard against it, we propose two alternative and robust inference methods: (i) explicit bias correction and bias-corrected confidence intervals; and (ii) joint maximum likelihood estimation. In a series of simulations and applications, we show that the two-step strategy produces material biases whereas both proposed methods perform well.

References

Adams, R. B., Ragunathan, V., and Tumarkin, R. (2021). Death by committee? An analysis of corporate board (sub-) committees. *Journal of Financial Economics*, 141(3):1119–1146.

Ahrens, M., Ashwin, J., Calliess, J.-P., and Nguyen, V. (2021). Bayesian Topic Regression for Causal Inference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8162–8188, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aigner, D. J. (1973). Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics*, 1(1):49–59.

Aksoy, C. G., Barrero, J. M., Bloom, N., Davis, S. J., Dolls, M., and Zarate, P. (2022). Working From Home Around the World.

Allon, G., Chen, D., Jiang, Z., and Zhang, D. (2023). Machine Learning and Prediction Errors in Causal Inference. *SSRN Electronic Journal*.

Arora, S., Ge, R., Kannan, R., and Moitra, A. (2012). Computing a nonnegative matrix factorization – provably. In *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing*, STOC '12, pages 145–162, New York, NY, USA. Association for Computing Machinery.

Ash, E., Morelli, M., and Vannoni, M. (2022). More Laws, More Growth? Evidence from U.S. States.

Avivi, H. (2024). Are Patent Examiners Gender Neutral? Unpublished manuscript.

Bai, J. and Ng, S. (2006). Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions. *Econometrica*, 74(4):1133–1150.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring Economic Policy Uncertainty*. *The Quarterly Journal of Economics*, 131(4):1593–1636.

Bandiera, O., Prat, A., Hansen, S., and Sadun, R. (2020). CEO Behavior and Firm Performance. *Journal of Political Economy*, 128(4):1325–1369.

Barrero, J. M., Bloom, N., and Davis, S. J. (2021). Why Working from Home Will Stick.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. *The Quarterly Journal of Economics*, 120(1):387–422.

Betancourt, M. (2018). A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]*.

Bing, X., Bunea, F., and Wegkamp, M. (2020). Optimal estimation of sparse topic models. *Journal of Machine Learning Research*, 21:1–45.

Blei, D. M. and McAuliffe, J. D. (2010). Supervised Topic Models. *arXiv:1003.0783 [stat]*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(null):993–1022.

Boxell, L. and Conway, J. (2022). Journalist Ideology and the Production of News: Evidence from Movers. *SSRN Electronic Journal*.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. (2018). JAX: Composable transformations of Python+NumPy programs.

Bybee, L., Kelly, B., Manela, A., and Xiu, D. (2024). Business News and Business Cycles. *The Journal of Finance*, 79(5):3105–3147.

Chen, X., Christensen, T. M., and Tamer, E. (2018). Monte Carlo Confidence Sets for Identified Sets. *Econometrica*, 86(6):1965–2018.

Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric Efficiency in GMM Models with Auxiliary Data. *The Annals of Statistics*, 36(2):808–843.

Chesher, A. (1991). The effect of measurement error. *Biometrika*, 78(3):451–462.

Compiani, G., Morozov, I., and Seiler, S. (2023). Demand Estimation with Text and Image Data. Technical Report 10695, CESifo.

Draca, M. and Schwarz, C. (2021). How Polarized are Citizens? Measuring Ideology from the Ground-Up. SSRN Scholarly Paper ID 3154431, Social Science Research Network, Rochester, NY.

Egami, N., Hinck, M., Stewart, B., and Wei, H. (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.

Einav, L., Finkelstein, A., and Mahoney, N. (2022). Producing Health: Measuring Value Added of Nursing Homes.

Evdokimov, K. S. and Zeleneev, A. (2023). Simple Estimation of Semiparametric Models with Measurement Errors.

Fong, C. and Tyler, M. (2021). Machine Learning Predictions as Regression Covariates.

Political Analysis, 29(4):467–484.

Freyaldenhoven, S., Ke, S., Li, D., and Olea, J. L. M. (2023). On the Testability of the Anchor Words Assumption in Topic Models.

Gabaix, X., Koijen, R. S. J., and Yogo, M. (2023). Asset Embeddings. *SSRN Electronic Journal*.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019). Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, 87(4):1307–1340.

Gonçalves, S. and Perron, B. (2014). Bootstrapping factor-augmented regression models. *Journal of Econometrics*, 182(1):156–173.

Gorodnichenko, Y., Pham, T., and Talavera, O. (2023). The Voice of Monetary Policy. *American Economic Review*, 113(2):548–584.

Hahn, J. and Kuersteiner, G. (2002). Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects when Both n and T Are Large. *Econometrica*, 70(4):1639–1657.

Hansen, S., Lambert, P. J., Bloom, N., Davis, S. J., Sadun, R., and Taska, B. (2023). Remote Work across Jobs, Companies, and Space.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*. *The Quarterly Journal of Economics*, 133(2):801–870.

Hoberg, G. and Phillips, G. (2016). Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*, 124(5):1423–1465.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47):1593–1623.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley California USA. ACM.

Ke, S., Olea, J. L. M., and Nesbit, J. (2021). Robust Machine Learning Algorithms for Text Analysis. Unpublished manuscript.

Ke, Z. T. and Wang, M. (2022). Using SVD for Topic Modeling. *Journal of the American Statistical Association*, pages 1–16.

Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021). Measuring Technological Innovation over the Long Run. *American Economic Review: Insights*, 3(3):303–320.

Lambert, P. J., Bloom, N., Davis, S., Hansen, S., Muvdi, Y., Sadun, R., and Taska, B. (2023). Research: The Growing Inequality of Who Gets to Work from Home. *Harvard Business Review*.

Larsen, V. H. and Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, 210(1):203–218.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, UK ; New York, illustrated edition edition.

Magnolfi, L., McClure, J., and Sorensen, A. (2022). Embeddings and Distance-based Demand for Differentiated Products. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, page 607, New York, NY, USA. Association for Computing Machinery.

Malmendier, U., Nagel, S., and Yan, Z. (2021). The making of hawks and doves. *Journal of Monetary Economics*, 117:19–42.

Mardia, J., Jiao, J., Tánczos, E., Nowak, R. D., and Weissman, T. (2019). Concentration Inequalities for the Empirical Distribution.

Mueller, H. and Rauh, C. (2018). Reading Between the Lines: Prediction of Political Violence Using Newspaper Text. *American Political Science Review*, 112(2):358–375.

Munro, E. and Ng, S. (2022). Latent Dirichlet Analysis of Categorical Survey Responses. *Journal of Business & Economic Statistics*, 40(1):256–271.

Neal, R. M. (2012). MCMC using Hamiltonian dynamics. *arXiv:1206.1901 [physics, stat]*.

Nimczik, J. S. (2017). Job Mobility Networks and Endogenous Labor Markets. Technical Report 168147, Verein für Socialpolitik / German Economic Association.

Olivella, S., Pratt, T., and Imai, K. (2021). Dynamic Stochastic Blockmodel Regression for Network Data: Application to International Militarized Conflicts. *arXiv:2103.00702 [cs, stat]*.

Pagan, A. (1984). Econometric Issues in the Analysis of Regressions with Generated Regressors. *International Economic Review*, 25(1):221–247.

Phan, D., Pradhan, N., and Jankowiak, M. (2019). Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv:1912.11554 [cs, stat]*.

Phillips, P. C. B. (1987). Towards a unified asymptotic theory for autoregression. *Biometrika*, 74(3):535–547.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4):1064–1082.

Ruiz, F. J. R., Athey, S., and Blei, D. M. (2020). SHOPPER: A probabilistic model of consumer choice with substitutes and complements. *The Annals of Applied Statistics*, 14(1):1–27.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*.

Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3):557–586.

Stock, J. H. and Watson, M. W. (2002). Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Thorsrud, L. A. (2020). Words are the New Numbers: A Newsy Coincident Index of the Business Cycle. *Journal of Business & Economic Statistics*, 38(2):393–409.

Vafa, K., Athey, S., and Blei, D. M. (2023). Decomposing Changes in the Gender Wage Gap over Worker Careers. In *NBER Summer Institute*, Boston, MA.

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

Wu, R., Zhang, L., and Tony Cai, T. (2023). Sparse Topic Modeling: Computational Efficiency, Near-Optimal Algorithms, and Statistical Inference. *Journal of the American Statistical Association*, 118(543):1849–1861.

Zhang, J., Xue, W., Yu, Y., and Tan, Y. (2023). Debiasing Machine-Learning- or AI-Generated Regressors in Partial Linear Models. *SSRN Electronic Journal*.

A Proofs of Main Results

Here we present proofs of just the main results in the text. Proofs of additional results are deferred to Appendix F.

Proof of Theorem 1. First consider the denominator. We have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T + \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\xi}}_i-\boldsymbol{\xi}_i)(\hat{\boldsymbol{\xi}}_i-\boldsymbol{\xi}_i)^T + \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\xi}}_i-\boldsymbol{\xi}_i)\boldsymbol{\xi}_i^T + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}_i(\hat{\boldsymbol{\xi}}_i-\boldsymbol{\xi}_i)^T \\ &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T + \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix} \\ &\quad + \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\boldsymbol{\theta}_i^T & \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\mathbf{q}_i^T \\ \mathbf{0} & \mathbf{0}\end{bmatrix} \\ &\quad + \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\theta}_i(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)^T & \mathbf{0} \\ \frac{1}{n}\sum_{i=1}^{n}\mathbf{q}_i(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)^T & \mathbf{0}\end{bmatrix}.\end{aligned}$$

Hence,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T \to_p \mathbb{E}\Big[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\Big] + \begin{bmatrix}\mathbf{V}+\mathbf{W}+\mathbf{W}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}$$

by Assumption 1(ii). The right-hand side is finite by Assumption 1(i) and invertible by assumption.

For the numerator term, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_iY_i &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\boldsymbol{\psi} + \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\xi}}_i-\boldsymbol{\xi}_i)\boldsymbol{\xi}_i^T\boldsymbol{\psi} + \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\varepsilon_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\boldsymbol{\psi} + \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\boldsymbol{\theta}_i^T & \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\mathbf{q}_i^T \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\boldsymbol{\psi} + \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\varepsilon_i.\end{aligned}$$

Hence,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_iY_i \to_p \left(\mathbb{E}\Big[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\Big] + \begin{bmatrix}\mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right)\boldsymbol{\psi}$$

by Assumption 1(ii). The first result now follows by Slutsky's theorem. The second result then follows because $(\mathbf{A}+\mathbf{Q})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{Q}\mathbf{A}^{-1} + O(\|\mathbf{Q}\|^2)$ for $\mathbf{A}$ invertible and $\mathbf{Q}$ small. ∎

Proof of Theorem 2. First consider the denominator. We have $\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T \to_p \mathbb{E}\Big[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\Big]$ by Assumption 2(ii). Result (11) now follows by Assumptions 2(i) and 2(iii) and Sltusky's theorem.

To establish (10), first write

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\psi}}-\boldsymbol{\psi}) &= \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i(Y_i-\hat{\boldsymbol{\xi}}_i^T\boldsymbol{\psi})\right)\\ &= \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}_i)^T\boldsymbol{\gamma}\right)+\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\varepsilon_i\right)\\ &=: T_{1,n}+T_{2,n}.\end{aligned}$$

It follows by Assumption 2(i)-(iii) and the Continuous Mapping Theorem that

$$T_{2,n}\to_d N\left(\mathbf{0},\,\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]^{-1}\mathbb{E}\left[\varepsilon_i^2\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]^{-1}\right).$$

For the remaining term, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i(\boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}_i)^T\boldsymbol{\gamma}=\begin{bmatrix}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)^T\boldsymbol{\gamma}\\ -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{q}_i(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)^T\boldsymbol{\gamma}\end{bmatrix}\to_p\begin{bmatrix}-\kappa\boldsymbol{\Omega}\boldsymbol{\gamma}\\ \mathbf{0}\end{bmatrix}$$

by Assumption 2(ii). Hence,

$$T_{1,n}\to_p-\kappa\,\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]^{-1}\begin{bmatrix}\boldsymbol{\Omega}\boldsymbol{\gamma}\\ \mathbf{0}\end{bmatrix}$$

by Assumptions 2(i) and 2(ii) and Slutsky's theorem. ∎

Proof of Theorem 3. First note that

$$\Pr(\psi_j\in\mathrm{CI}(\psi_j))=\Pr\left(\left|\sqrt{n}\left(\hat{\psi}_j-\psi_j\right)-\hat{b}_j\right|\le z_{1-\alpha/2}\,\hat{\sigma}_j\right),$$

where $\hat{\sigma}_j\to_p\sigma_j$ by Theorem 2, with σ_j the square root of the jth diagonal entry of the asymptotic variance on the right-hand side of (11). Moreover, σ_j is positive and finite by the rank condition in the statement of the theorem and Assumption 2(i). The proof of Theorem 2 shows that $\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1}\to_p\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]^{-1}$ and Theorem 2 itself implies $\hat{\boldsymbol{\gamma}}\to_p\boldsymbol{\gamma}$. Hence, by consistency of $\hat{\kappa}$ and $\hat{\boldsymbol{\Omega}}$, we have that $\hat{\boldsymbol{b}}\to_p\boldsymbol{b}$. The result follows by (10) and the continuous mapping theorem. ∎

Proof of Lemma 1. It is enough to show $\widehat{FPR}/\mathbb{E}[FP_i]\to_p 1$. Conditional on $(\mathbf{x}_i,\mathbf{q}_i)_{i=1}^m$, each $\hat{\theta}_i(1-\theta_i)$ are independent Bernoulli random variables with success probability FP_i. By Chernoff's inequality, for any $\delta>0$ we have

$$\Pr\left(\left|\frac{\widehat{FPR}}{\mathbb{E}[FP_i]}-1\right|>\delta\,\middle|\,(\mathbf{x}_i,\mathbf{q}_i)_{i=1}^m\right)\le 2e^{-\delta^2\sum_{i=1}^m FP_i/3}. \tag{14}$$

By Chebyshev's inequality, with probability at least $1 - C^{-2}$ we have

$$\sum_{i=1}^{m} FP_i \geq m\,\mathbb{E}[FP_i] - C\sqrt{m\,\mathbb{E}[FP_i^2]},$$

for any $C > 0$. But $\mathbb{E}[FP_i^2] < \mathbb{E}[FP_i]$ because $0 \leq FP_i \leq 1$. The conditions $\sqrt{n}\,\mathbb{E}[FP_i] \to \kappa > 0$ and $n/m^2 \to 0$ together imply that $m\,\mathbb{E}[FP_i] \to +\infty$. Hence we can take $C \to +\infty$ with $C/\sqrt{m\,\mathbb{E}[FP_i]} \to 0$ to deduce that

$$\sum_{i=1}^{m} FP_i \geq \frac{1}{2} m\,\mathbb{E}[FP_i]$$

holds with probability approaching one. Letting A_n denote the sequence of sets of $(\mathbf{x}_i, \mathbf{q}_i)_{i=1}^m$ upon which the preceding inequality holds, we have

$$\begin{aligned}
\Pr\left(\left|\frac{\widehat{FPR}}{\mathbb{E}[FP_i]} - 1\right| > \delta\right) &\leq \mathbb{E}\left[\Pr\left(\left|\frac{\widehat{FPR}}{\mathbb{E}[FP_i]} - 1\right| > \delta \,\middle|\, (\mathbf{x}_i, \mathbf{q}_i)_{i=1}^m\right) \mathbb{I}[(\mathbf{x}_i, \mathbf{q}_i)_{i=1}^m \in A_n)\right] \\
&\quad + \Pr((\mathbf{x}_i, \mathbf{q}_i)_{i=1}^m \in A_n^c) \\
&\leq 2e^{-\delta^2 m\,\mathbb{E}[FP_i]/6} + \Pr((\mathbf{x}_i, \mathbf{q}_i)_{i=1}^m \in A_n^c) \to 0,
\end{aligned}$$

by (14) and the fact that $m\,\mathbb{E}[FP_i] \to +\infty$. ∎

Proof of Lemma 2. Consistency of $\hat{\boldsymbol{\Omega}}$ follows by similar arguments to Lemmas 5 and 6, using the condition $\bar{\boldsymbol{w}}_n \to_p \mathbb{E}[\boldsymbol{w}_i]$. Moreover, by Chebyshev's inequality, for any $\delta > 0$ we have

$$\Pr\left(\left|\hat{\kappa} - \sqrt{n}\,\mathbb{E}[C_i^{-1}]\right| > \delta\right) \leq \frac{1}{\delta^2}\mathbb{E}[C_i^{-2}].$$

As $C_i \geq 1$ and $\sqrt{n}\,\mathbb{E}[C_i^{-1}] \to \kappa \geq 0$, we have $\mathbb{E}[C_i^{-2}] \leq \mathbb{E}[C_i^{-1}] \to 0$. Hence, $\hat{\kappa} \to_p \kappa$. ∎

B Applications

In this section we develop theory for the two-step strategy when (i) and (ii) a topic model are deployed as the upstream information retrieval model.

B.1 AI/ML-Generated Labels

Here we first generalize the basic framework from Section 3.1.1 to allow for multiple categories. Let the vector $\boldsymbol{\theta}_i$ indicate membership of K distinct categories. Thus, if individual i belongs to category k, we have $\theta_{i,k} = 1$ and $\theta_{i,j} = 0$ for all $j \neq k$. Let $p_k(\mathbf{x}_i)$ denote the true conditional probability $\Pr(\theta_{i,k} = 1|\mathbf{x}_i)$, and let $\mathbf{p}(\mathbf{x}_i) = (p_k(\mathbf{x}_i))_{k=1}^K$. Similarly, let $\pi_k(\mathbf{x}_i)$ denote the probability with which the classifier assigns label k given $\mathbf{x}_i$, and let $\boldsymbol{\pi}(\mathbf{x}_i) = (\pi_k(\mathbf{x}_i))_{k=1}^K$. To simplify notation, we write these probabilities as a function of $\mathbf{x}_i$ only with the understanding that $\mathbf{q}_i$ is included as a component of $\mathbf{x}_i$ if it is relevant for predicting $\boldsymbol{\theta}_i$. We introduce a function $\mathbf{r}(\mathbf{x}_i, \cdot) : [0,1] \to \{0,1\}^K$ and, for each observation i, a random variable $U_i \sim U[0,1]$ drawn independent of $(\mathbf{x}_i, \mathbf{q}_i, Y_i, \boldsymbol{\theta}_i)$ and all other U_j, $j \neq i$, so that $\mathbf{r}(\mathbf{x}_i, U_i)|\mathbf{x}_i \sim \text{Multinomial}(1, \boldsymbol{\pi}(\mathbf{x}_i))$. For instance, if $K = 1$ we could have $\mathbf{r}(\mathbf{x}_i, U_i) = \mathbb{I}[U_i \leq \pi_1(\mathbf{x}_i)]$ in the case of randomized predictions, and $\mathbf{r}(\mathbf{x}_i, U_i) = \mathbb{I}[\pi_1(\mathbf{x}_i) \geq \frac{1}{2}]$ in the case of Bayes classifiers.

To simplify some expressions, we further assume that $\mathbb{E}[\varepsilon_i|\mathbf{x}_i] = 0$. Thus, the unstructured data $\mathbf{x}_i$ contains no additional information for predicting ε_i beyond that contained in the group label $\boldsymbol{\theta}_i$ and $\mathbf{q}_i$.

B.1.1 Fixed Population

We first provide primitive sufficient conditions for the theoretical results for the fixed-population case. The following assumptions are required to hold for a fixed distribution of $(Y_i, \mathbf{q}_i, \mathbf{x}_i, \boldsymbol{\theta}_i)_{i=1}^n$ as the sample size n becomes large.

Assumption 3. *(i)* $\max_{1 \leq i \leq n} \|\hat{\boldsymbol{\theta}}_i - \mathbf{r}(\mathbf{x}_i, U_i)\| \to_p 0$.

(ii) $\mathbb{E}\left[\|\mathbf{q}_i\|^2\right] < \infty$, $\mathbb{E}\left[(1 + \|\mathbf{q}_i\|)|\varepsilon_i|\right] < \infty$, *and* $\mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]$ *has full rank.*

(iii) $\mathbb{E}\left[(\boldsymbol{\pi}(\mathbf{x}_i) - \mathbf{p}(\mathbf{x}_i))\, \mathbf{q}_i^T\right] = \mathbf{0}$.

Assumption 3(i) imposes some structure on the AI/ML-generated predictions $\hat{\boldsymbol{\theta}}_i$. It allows for the classification algorithm to be pre-trained, provided the pre-training error is of smaller asymptotic order. For some intuition, consider the case of randomized predictions with $K = 1$ and let $\hat{\boldsymbol{\theta}}_i = \mathbb{I}[U_i \leq \pi_1(\mathbf{x}_i; \hat{\boldsymbol{\mu}})]$ and $\mathbf{r}(\mathbf{x}_i, U_i) = \mathbb{I}[U_i \leq \pi_1(\mathbf{x}_i; \boldsymbol{\mu})]$ where $\hat{\boldsymbol{\mu}}$ is an estimator of $\boldsymbol{\mu}$ and the U_i are independent $U[0,1]$. Note that $\hat{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{\theta}}_j$ are statistically dependent because of their shared dependence on $\hat{\boldsymbol{\mu}}$, whereas $\mathbf{r}(\mathbf{x}_i, U_i)$

and $\mathbf{r}(\mathbf{x}_j, U_j)$ are independent. But $\Pr(\hat{\boldsymbol{\theta}}_i \neq \mathbf{r}(\mathbf{x}_i, U_i)|\mathbf{x}_i, \hat{\boldsymbol{\mu}}) \leq |\pi_1(\mathbf{x}_i; \hat{\boldsymbol{\mu}}) - \pi_1(\mathbf{x}_i; \boldsymbol{\mu})|$, so Assumption 3(i) holds under appropriate convergence conditions on $\hat{\boldsymbol{\mu}}$.

Assumption 3(ii) imposes standard moment and rank conditions. Finally, Assumption 3(iii) imposes an "unbiasedness" or "fairness" condition on the classifier. If there are no covariates so that $\mathbf{q}_i$ contains only a constant or if $\mathbf{q}_i$ and $(\boldsymbol{\theta}_i, \hat{\boldsymbol{\theta}}_i)$ are independent, then this assumption says that the classifier assigns individuals to the K categories with the true probabilities (unconditionally). Note this condition is much weaker than requiring conditional unbiasedness: $\boldsymbol{\pi}(\mathbf{x}_i) = \mathbf{p}(\mathbf{x}_i)$ for all $\mathbf{x}_i$. Moreover, in the case of logistic classifiers it is an implication of the first-order condition for fitting parameters when $\mathbf{q}_i$ is included as a component of $\mathbf{x}_i$.

Our first result for AI/ML-generated labels shows that the OLS estimator $\hat{\boldsymbol{\psi}}$ is inconsistent and characterizes its bias analytically.

Theorem 4. *Suppose that Assumption 3 holds. Then Assumption 1 holds and the OLS estimator $\hat{\boldsymbol{\psi}}$ has probability limit given by (7) with*

$$\begin{aligned}\mathbf{V} &= \mathbb{E}\left[\operatorname{diag}(\boldsymbol{\pi}(\mathbf{x}_i) + \mathbf{p}(\mathbf{x}_i)) - \boldsymbol{\pi}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^T - \mathbf{p}(\mathbf{x}_i)\boldsymbol{\pi}(\mathbf{x}_i)^T\right], \\ \mathbf{W} &= \mathbb{E}\left[\boldsymbol{\pi}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^T - \operatorname{diag}(\mathbf{p}(\mathbf{x}_i))\right].\end{aligned}$$

If $\mathbf{V}$ and $\mathbf{W}$ are small, then first-order bias is proportional to

$$\mathbf{V} + \mathbf{W}^T = \mathbb{E}\left[\operatorname{diag}(\boldsymbol{\pi}(\mathbf{x}_i)) - \boldsymbol{\pi}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^T\right].$$

B.1.2 Sequence of Populations

In modern settings where high-performance classifiers are deployed at scale, there may be both a relatively large sample size n and a relatively low degree of misclassification. To mimic this setting, we consider a sequence of populations where for each n the distribution of $(Y_i, \boldsymbol{\theta}_i, \mathbf{q}_i)$ is fixed and satisfies (4) but the distribution of $\mathbf{x}_i|(Y_i, \boldsymbol{\theta}_i, \mathbf{q}_i)$ varies with n so that $\mathbf{x}_i$ becomes more informative about $\boldsymbol{\theta}_i$ and misclassification errors become small. To this end, consider the matrix $\mathbf{V}$ from Theorem 4. Misclassification rates for each of the K labels are collected along the diagonal of $\mathbf{V}$:

$$(\mathbf{V})_{k,k} = \mathbb{E}\left[\pi_k(\mathbf{x}_i) + p_k(\mathbf{x}_i) - 2\pi_k(\mathbf{x}_i)p_k(\mathbf{x}_i)\right].$$

The first condition we require is that the sum of the misclassification rates vanishes:

$$\operatorname{tr}(\mathbf{V}) \to 0 \tag{15}$$

as $n \to \infty$. This condition requires that the true probabilities $p_k(\mathbf{x}_i) = \Pr(\theta_{i,k} = 1|\mathbf{x}_i)$ converge to zero or one (i.e., accurate prediction of $\boldsymbol{\theta}_i$ is possible given $\mathbf{x}_i$), and that the

differences $\pi_k(\mathbf{x}_i) - p_k(\mathbf{x}_i)$ converge to zero (i.e., the classifier produces correct labels).

We also place some structure on the false-positive rates. Let $FP_i = \sum_{k=1}^{K} FP_k(\mathbf{x}_i)$ denote the total false-positive rate for individual i, where $FP_k(\mathbf{x}_i) = \pi_k(\mathbf{x}_i)(1 - p_k(\mathbf{x}_i))$ denotes the individual's false-positive probability for label k. We require

$$\sqrt{n}\,\mathbb{E}\left[\text{diag}(\boldsymbol{\pi}(\mathbf{x}_i)) - \boldsymbol{\pi}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^T\right] \to \kappa\,\boldsymbol{\Omega}, \tag{16}$$

where

$$\lim_{n\to\infty} \sqrt{n}\,\mathbb{E}\left[FP_i\right] = \kappa \geq 0,$$

and

$$\lim_{n\to\infty} \frac{\mathbb{E}\left[\text{diag}(\boldsymbol{\pi}(\mathbf{x}_i)) - \boldsymbol{\pi}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^T\right]}{\mathbb{E}\left[FP_i\right]} = \boldsymbol{\Omega}, \tag{17}$$

assuming both limits exist. In words, $\kappa = 0$ corresponds to a case where the false-positive rate across all categories vanishes faster than sampling error. Conversely, κ positive allows for the total false-positive rate to be of the same order as sampling error. If there is a single category ($K = 1$) then $\boldsymbol{\Omega} = 1$. More generally, $\boldsymbol{\Omega}$ quantifies the relative frequency with which false-positives occur among the K alternatives.

In what follows, notions of convergence in probability and in distribution should be understood as holding along a sequence of populations satisfying the above conditions.

Assumption 4. *(i)* $\sqrt{n}\max_{1\leq i\leq n}\|\hat{\boldsymbol{\theta}}_i - \mathbf{r}(\mathbf{x}_i, U_i)\| \to_p 0$.

(ii) $\mathbb{E}\left[\|\mathbf{q}_i\|^4\right] < \infty$, $\mathbb{E}\left[\varepsilon_i^4\right] < \infty$, *and* $\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]$ *has full rank.*

(iii) $\mathbb{E}\left[\left(\boldsymbol{\pi}(\mathbf{x}_i) - \mathbf{p}(\mathbf{x}_i)\right)\mathbf{q}_i^T\right] = \mathbf{0}$.

Assumption 4(i) strengthens Assumption 3(i) to require convergence at a faster-than-root-n rate. Assumption 4(ii) is standard. Assumption 4(iii) is the same as Assumption 3(iii).

Our second main result derives the asymptotic distribution of $\hat{\boldsymbol{\psi}}$, characterizes its bias analytically, and establishes consistency of two-step standard errors.

Theorem 5. *Suppose that Assumption 4 holds. Then Assumption 2 holds and the OLS estimator* $\hat{\boldsymbol{\psi}}$ *has asymptotic distribution given by (10) with* $\boldsymbol{\Omega}$ *as given in (17). Moreover, two-step standard errors are consistent.*

B.2 Topic Models

The framework is introduced in Section 3.1.2. Each unstructured observation is a V-dimensional vector $\mathbf{x}_i$ of feature counts. The multinomial probabilities $\mathbf{p}_i = \mathbb{E}[\mathbf{x}_i/C_i]$ with $C_i = \sum_{v=1}^{V} x_{i,v}$ have a factor structure $\mathbf{p}_i = \mathbf{B}^T\boldsymbol{w}_i$. The columns of the $V \times K$

matrix $\mathbf{B}^T$ are the $K \ll V$ common factors ("topics") and $\boldsymbol{w}_i$ is an observation-specific distribution over topics. In both asymptotic frameworks below, the data are a random sample $(Y_i, \mathbf{q}_i, \mathbf{x}_i, C_i)_{i=1}^n$ satisfying the regression equation (4) with $\boldsymbol{\theta}_i = \boldsymbol{S}\boldsymbol{w}_i$ for a known selection matrix $\boldsymbol{S}$ and the multinomial-factor distribution (5).

To simplify some expressions, we further assume that the document size C_i is independent of $(\boldsymbol{w}_i, \mathbf{q}_i, Y_i)$. We also assume that $\mathbf{x}_i$ and $\mathbf{q}_i$ are independent conditional on $(C_i, \boldsymbol{w}_i)$, and that ε_i and $(\mathbf{x}_i, C_i)$ are independent conditional on $(\boldsymbol{w}_i, \mathbf{q}_i)$. In effect, the latter two assumptions ensure the multinomial noise $(\hat{\mathbf{p}}_i - \mathbf{p}_i)$ is uncorrelated with $(\varepsilon_i, \mathbf{q}_i, \boldsymbol{\theta}_i)$ while allowing for correlation between $\mathbf{q}_i$ and $\boldsymbol{\theta}_i$. These assumptions seem very reasonable and can be relaxed: doing so simply complicates the expressions below. We also slightly strengthen (4) to require that $\mathbb{E}[\varepsilon_i | \boldsymbol{w}_i, \mathbf{q}_i] = 0$. That is, the conditional mean of Y_i given $(\boldsymbol{w}_i, \mathbf{q}_i)$ depends only on $\mathbf{q}_i$ and the included topic weights $\boldsymbol{\theta}_i = \boldsymbol{S}\boldsymbol{w}_i$.

Remark 4 (Identification)**.** We also assume that $\mathbf{B}$ is identified. That is, there is a unique $K \times V$ row-stochastic matrix $\mathbf{B}$ such that the mapping $(\mathbf{B}, \boldsymbol{w}) \mapsto \mathbf{p} = \mathbf{B}^T\boldsymbol{w}$ is invertible on $(\Delta^{V-1})^K \times \mathcal{W}$ with $(\Delta^{V-1})^K$ denoting the space of $K \times V$ row-stochastic matrices and $\mathcal{W}$ denoting the support of $\boldsymbol{w}_i$. This notion of identification is appropriate because in our framework the latent $\boldsymbol{w}_i$ are observation-specific and the number of observations n becomes large. We emphasize that this is different, and much weaker, than the notion of identification studied to date in the econometrics literature (e.g., Ke et al. (2021), Freyaldenhoven et al. (2023)), which treats n as fixed and considers uniqueness of the factorization $\mathbf{P} = \mathbf{B}^T\boldsymbol{W}$ with $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_n]$ and $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n]$ conditional on the n sampled units. Within a fixed-n context, uniqueness is commonly achieved in text applications by assuming the existence of anchor words that are known to appear in some topics but not others. Anchor words may also be used to guarantee identification in our setting, but they are not necessary when $\boldsymbol{w}_i$ has a rich support.

There are many different ways of estimating $\mathbf{B}$ and $\boldsymbol{w}_i$ in (5). For instance, one could use LDA (Blei et al. 2003) or more recent methods developed by Bing et al. (2020), Wu et al. (2023), Ke and Wang (2022), and many others. As our objective is to focus on the consequences of the two-step strategy, we abstract from algorithmic-specific details and instead impose mild conditions on the estimators $\hat{\mathbf{B}}$ of $\mathbf{B}$ and $\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_n$ of $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$. Let $\hat{\mathbf{p}}_i = \mathbf{x}_i / C_i$, $i = 1, \ldots, n$.

B.2.1 Fixed Population

We first provide primitive sufficient conditions for the theoretical results for the fixed-population case. The following assumptions are required to hold for a fixed distribution of $(Y_i, \mathbf{q}_i, \mathbf{x}_i, \boldsymbol{w}_i, C_i)_{i=1}^n$ as the sample size n becomes large.

Assumption 5. *(i)* $\mathbf{B}$ *has full rank.*

(ii) $\hat{\mathbf{B}} \to_p \mathbf{B}$.

(iii) $\max_{1\leq i\leq n} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\hat{\mathbf{p}}_i\| \to_p 0$.

(iv) $\mathbb{E}\left[\|\mathbf{q}_i\|^2\right] < \infty$, $\mathbb{E}\left[(1+\|\mathbf{q}_i\|)|\varepsilon_i|\right] < \infty$, *and* $\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]$ *has full rank.*

Assumption 5(i) says that there are at least K topics. This is a weak restriction, as $K \ll V$ in applications. Assumption 5(ii) says that $\hat{\mathbf{B}}$ is consistent for $\mathbf{B}$, which is satisfied by many estimators for topic models. Assumption 5(iii) imposes some structure on the estimators $\hat{\boldsymbol{w}}_i$. This condition is not vacuous: $\boldsymbol{w}_i = (\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\mathbf{p}_i$ by Assumption 5(i), so, given any consistent estimator $\hat{\mathbf{B}}$ of $\mathbf{B}$, one could set $\hat{\boldsymbol{\theta}}_i = \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\hat{\mathbf{p}}_i$. In that case, $\max_{1\leq i\leq n} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\hat{\mathbf{p}}_i\| = 0$. Assumption 5(i)-(iii) holds trivially for the pure multinomial model because $\hat{\mathbf{B}} = \mathbf{B} = \mathbf{I}$ and $\hat{\mathbf{p}}_i = \hat{\boldsymbol{w}}_i$. Finally, Assumption 5(iv) imposes standard moment conditions and ensures that the regressors are not perfectly collinear.

Our first result for topic models shows that the OLS estimator $\hat{\boldsymbol{\psi}}$ is inconsistent. We also characterize the bias analytically. Let $\text{diag}(\mathbf{v})$ denote a matrix with $\mathbf{v}$ down its diagonal and whose off-diagonal elements are all zero.

Theorem 6. *Suppose that Assumption 5 holds. Then Assumption 1 holds and the OLS estimator* $\hat{\boldsymbol{\psi}}$ *has probability limit given by (7) with*

$$\mathbf{V} = \mathbb{E}\left[\frac{1}{C_i}\right]\left(\boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\,\text{diag}(\mathbf{B}^T\mathbb{E}[\boldsymbol{w}_i])\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\boldsymbol{S}^T - \mathbb{E}\left[\boldsymbol{\theta}_i\boldsymbol{\theta}_i^T\right]\right), \quad \mathbf{W} = \mathbf{0}.$$

B.2.2 Sequence of Populations

To capture a more realistic setting where the amount of unstructured data per observation is large, we consider a sequence of populations in which for each n, the distribution of $(Y_i, \mathbf{q}_i, \boldsymbol{w}_i)$ fixed and satisfies (4), and the conditional distribution of $(\mathbf{x}_i, C_i)$ given $(Y_i, \mathbf{q}_i, \boldsymbol{w}_i)$ vary with n so that (5) holds and

$$\sqrt{n}\,\mathbb{E}\left[\frac{1}{C_i}\right] \to \kappa \geq 0. \tag{18}$$

In what follows, notions of convergence in probability and in distribution should be understood as holding along this sequence of populations.

Assumption 6. *(i)* $\mathbf{B}$ *has full rank.*

(ii) $\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B}) \to_p 0$.

(iii) $\sqrt{n}\max_{1\leq i\leq n} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\hat{\mathbf{p}}_i\| \to_p 0$.

(iv) $\mathbb{E}\left[\|\mathbf{q}_i\|^4\right] < \infty$, $\mathbb{E}\left[\varepsilon_i^4\right] < \infty$, *and* $\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]$ *has full rank.*

(v) $C_i \gtrsim (\log n)^{1+\epsilon}$ *almost surely for some* $\epsilon > 0$.

Assumption 6(i) is the same as Assumption 5(i). Assumption 6(ii)-(iii) strengthens Assumption 5(ii)-(iii) to require convergence at a faster-than-root-n rate. We believe Assumption 6(ii) is broadly satisfied in view of known convergence rates for estimators of $\mathbf{B}$. For instance, Bing et al. (2020), Wu et al. (2023), and Ke and Wang (2022) derive finite-sample guarantees for different estimators $\hat{\mathbf{B}}$ of $\mathbf{B}$. Each of their results implies the corresponding estimator $\hat{\mathbf{B}}$ converges at the optimal rate $(nC)^{-1/2}$ (up to log terms) where, for simplicity, the C_i are all of the same order C. Hence, all estimators $\hat{\mathbf{B}}$ converge faster than $n^{-1/2}$ when C grows with n, as we have here by (18). Assumption 6(iii) is made to simplify derivations but is not vacuous: given any estimator $\hat{\mathbf{B}}$, one could set $\hat{\boldsymbol{\theta}}_i = \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\hat{\mathbf{p}}_i$, in which case $\sqrt{n}\max_{1\le i\le n}\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\hat{\mathbf{p}}_i\| = 0$. As before, Assumption 6(i)-(iii) holds trivially for the pure multinomial model because $\hat{\mathbf{B}} = \mathbf{B} = \mathbf{I}$ and $\hat{\mathbf{p}}_i = \hat{\boldsymbol{w}}_i$. Assumption 6(iv) is standard. Assumption 6(v) is made to simplify technical derivations and can be relaxed. It implies that C_i is supported on $[c(\log n)^{1+\epsilon}, \infty)$ for some $c, \epsilon > 0$. This is weaker than the conventional assumption that all C_i grow at the same rate C (Bing et al. 2020, Wu et al. 2023, Ke and Wang 2022) which, in view of (18), would imply that C_i is supported on $[cn^{1/2}, \infty)$ for some $c > 0$. This condition is only used to establish consistency of standard errors.

Our second main result for topic models derives the asymptotic distribution of $\hat{\boldsymbol{\psi}}$, characterizes its bias analytically, and establishes consistency of two-step standard errors.

Theorem 7. *Suppose that Assumption 6 holds. Then Assumption 2 holds and the OLS estimator* $\hat{\boldsymbol{\psi}}$ *has asymptotic distribution given by (10) with*

$$\boldsymbol{\Omega} = \boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\,\mathrm{diag}(\mathbf{B}^T\mathbb{E}[\boldsymbol{w}_i])\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\boldsymbol{S}^T - \mathbb{E}\left[\boldsymbol{\theta}_i\boldsymbol{\theta}_i^T\right].$$

Moreover, two-step standard errors are consistent.

C Further Details on Section 5.2.1

Table C.1 presents the parameters used for simulation exercise in Section 5.2.1. To ensure the model is properly identified, in each simulation we set $A = 100$ features to be 'anchor words' meaning that $\beta_{j,0}$ or $\beta_{j,1}$ is set to 0. We simulated data 200 times for each set and then estimated the model using 1-step approach, 2-step approach and the infeasible 2-step approach with known θ. We construct 95% confidence intervals for γ and ϕ using the corresponding 95% posterior credible intervals for these parameters. This construction is justified in view of the discussion in Section 4.2.1.

Table C.1: Parameters for the Simulation Exercise.

Parameter	Value	Description
(a) Data Simulation		
n	10000	Number of observations
V	300	Number of distinct features
C_i	$\{10, 25, 200\}$	Total number of features per document
K	2	Number of latent types
True ϕ	**1**	Effect of a covariates on un-normalized type shares
True γ	**5**	Effect of topic shares on numerical outcomes
True α	$(0, 1, 1, 1)$	Effect of additional covariates on numerical outcomes
g_i	$\sim N(0, \frac{\log(3)}{1.96})$	Covariate affecting type shares
$q_{i,m}$ $\forall m \in (1, 2, 3)$	$\sim N(0, 3)$	Additional covariates affecting outcome
σ_Y^2	16	SD of the numeric outcome's residual
σ_θ^2	1	SD of residual of the un-normalized type shares
η	0.2	Dirichlet concentration parameter
(b) Hyperparameters		
K	*as above*	Number of latent types
η	*as above*	Dirichlet concentration parameter
σ_θ^2	*as above*	SD of residual of the un-normalized type shares
$p(\phi_1)$	$N(0, 4)$	Prior for ϕ_1, i.e. $\sigma_\phi^2 = 4$
$p(\gamma_1)$	$N(0, 100)$	Prior for γ_1, i.e. $\sigma_\gamma^2 = 100$
$p(\alpha)$ $\forall m \in (0, 1, 2, 3)$	$N(0, 100)$	Prior for α, i.e. $\sigma_\alpha^2 = 100$
$p(\sigma_Y)$	Gamma$(1, 10)$	Prior for σ_Y, i.e. $s_0 = 1$ and $s_1 = 10$

We performed the simulation on a 'N1-highmem-2' instance on the Google Could Platform. The instance has 2 vCPUs and 13 GB of memory. We also utilized a single Tesla V100 GPU. We run chose 500 warmup and 500 post-warmup iterations. A single simulation (consisting of drawing the data, and estimating the model in three ways) took approximately 6 minutes.

D Example Code

```
from numpyro import sample, plate
import numpyro.distributions as dist
import jax.numpy as jnp
from jax.nn import softmax

class SUPTMC:
    def __init__(self, K, N, V, z, q, eta = .1, alpha = 1):
        self.K = K  # number of latent types
        self.N = N  # number of observations
        self.V = V  # number of distinct features
        self.z = z  # number of covariates affecting outcome
        self.q = q  # number of covariates affecting type shares
        self.eta = eta
        self.alpha = alpha

    def model(self, C, Z, Q, Y=None, X=None):
        # Supervised topic model with covariates

        # Y : regression outcome
        # X : feature count matrix
        # C : total number of features per observation
        # Z : covariates entering regression
        # Q : covariates entering type shares
        # K : number of types
        # eta, alpha : Dirichlet hyperparamters

        #### Upstream Factor Model ###

        with plate("topics", self.K):
            beta = sample("beta", dist.Dirichlet(
                self.eta * jnp.ones(self.V - self.num_anchors_per_class)))

        phis = sample("phis", dist.Normal(0,2).expand([self.q, self.K-1]))

        with plate_stack("docs", sizes = [self.N, self.K - 1]):
            A = sample("A", dist.Normal(jnp.matmul(Q, phis) , self.alpha))

        # document-topic distributions
        theta = deterministic(
            "theta",
            softmax(jnp.hstack([A, jnp.zeros([self.D, 1])]), axis = -1)
            )

        distMultinomial = dist.Multinomial(
            total_count=C,
            probs = jnp.matmul(theta, beta)
            )
        with plate("hist", self.N):
            X_bows = sample("obs_x", distMultinomial, obs = X)

        #### Downstream Regression Model ###

        gammas = sample("gammas", dist.Normal(0, 10).expand([self.K-1]))
        zetas = sample("zetas", dist.Normal(0, 10).expand([self.z]))
        sigma = sample("sigma", dist.Gamma(1, 10))

        mean = jnp.matmul(theta[:,:(self.K-1)], gammas) + jnp.matmul(Z, zetas)

        with plate("y", self.N):
            Y = sample("obs_y",  dist.Normal(mean, sigma), obs = Y)
```

Figure D.1: Numpyro's code used to estimate Supervised Topic Model with Covariates

Online Appendix: Inference for Regression with Variables Generated by AI or Machine Learning

Laura Battaglia
Oxford

Timothy Christensen
Yale

Stephen Hansen
UCL, IFS, and CEPR

Szymon Sacher
Meta

December 11, 2024

E Extensions and Complements

E.1 VARs

Several prominent studies, including Baker et al. (2016), have used text-derived measures as variables in vector autoregressions (VARs). These studies use a two-step strategy: variables of interest $\boldsymbol{\theta}_t$ (e.g., policy uncertainty) are estimated from unstructured data, then their estimates $\hat{\boldsymbol{\theta}}_t$ used as variables in a VAR alongside other variables $\mathbf{q}_t$ (e.g., interest rate, industrial production, unemployment). Standard inference on VAR parameters and impulse response functions (IRFs) is performed, treating $\hat{\boldsymbol{\theta}}_t$ as regular data.

Consider the VAR

$$\boldsymbol{\xi}_t = \boldsymbol{\psi}_0 + \boldsymbol{\Psi}_1 \boldsymbol{\xi}_{t-1} + \ldots + \boldsymbol{\Psi}_p \boldsymbol{\xi}_{t-p} + \varepsilon_t,$$

where $\boldsymbol{\xi}_t = (\boldsymbol{\theta}_t^T, \mathbf{q}_t^T)^T$. Since $\boldsymbol{\theta}_t$ is not observed, the VAR parameters $\boldsymbol{\psi} = (\boldsymbol{\psi}_0, \boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_p)$ are estimated by regressing $\hat{\boldsymbol{\xi}}_t = (\hat{\boldsymbol{\theta}}_t^T, \boldsymbol{q}_t^T)^T$ on its lagged values and a constant. Let $\hat{\boldsymbol{\psi}}$ denote the OLS estimator.

The theory developed above carries over to the VAR setting under suitable modifications of Assumptions 1 and 2. Partition the VAR coefficient matrices into blocks corresponding to whether (a) $\boldsymbol{\theta}_t$ or (b) $\mathbf{q}_t$ is the dependent variable. OLS estimators of the block (b) parameters will behave similarly to Theorems 1 and Theorem 2. That is, they are inconsistent in a fixed-population setting and $\sqrt{n}$-consistent and asymptotically normal with a location shift proportional to κ in a sequence-of-populations framework. OLS estimators of block (a) parameters have a measurement error $\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t$ in the dependent

variable. Under suitable regularity conditions this measurement error is asymptotically negligible in a sequence-of-populations framework. Hence, OLS estimators of block (a) parameters are $\sqrt{n}$-consistent and asymptotically normal with a location shift proportional to κ. Further, the asymptotic variance of $\hat{\boldsymbol{\psi}}$ will be the same as if the VAR was estimated on the true $\boldsymbol{\theta}_t$ and can be consistently estimated when $\kappa > 0$ (cf. Theorem 2).

These consequences carry over to inference on IRFs. VAR-based IRF estimators computed from $\hat{\boldsymbol{\psi}}$ will be $\sqrt{n}$-consistent and asymptotically normal with a location shift proportional to κ, so delta-method confidence intervals for IRFs will under-cover. Other regression-based estimators of IRFs, such as local projections, using $\hat{\boldsymbol{\xi}}_t$ will similarly suffer from a location shift that leads to confidence intervals that under-cover.

E.2 Similarity Measures

Another leading use case of unstructured data are regressions on similarity measures derived from term frequencies or topic weights. In this case, the two-step strategy first estimates similarity from unstructured data then regresses numerical outcomes on the estimated similarity measures. We show that this strategy leads to biased inference on regression parameters unless sampling error dominates measurement error.

Suppose that the data are a random sample $(Y_i, \mathbf{x}_{1,i}, \mathbf{x}_{2,i}, C_{1,i}, C_{2,i})_{i=1}^n$, where

$$\mathbf{x}_{t,i} \sim \text{Multinomial}(C_{t,i}, \mathbf{p}_{t,i}), \qquad t = 1, 2.$$

Each feature count vector $\mathbf{x}_{t,i}$ is a noisy signal of the true (latent) frequency $\mathbf{p}_{t,i}$. We are interested in performing inference on the parameter γ_1 in the regression model

$$Y_i = \gamma_0 + \gamma_1 \left(\mathbf{p}_{1,i} \cdot \mathbf{p}_{2,i}\right) + \varepsilon_i. \tag{19}$$

Consider a setting based on Kelly et al. (2021): Y_i denotes citations of patent i after it is filed, $\mathbf{x}_{1,i}$ is a vector of feature counts for patent i, and $\mathbf{x}_{2,i}$ is a vector of feature counts of an existing stock of patents at the time patent i was filed. The counts $\mathbf{x}_{1,i}$ and $\mathbf{x}_{2,i}$ are noisy signals of the true information contents $\mathbf{p}_{1,i}$ and $\mathbf{p}_{2,i}$ of the new patent and existing stock. The dissimilarity between $\mathbf{p}_{1,i}$ and $\mathbf{p}_{2,i}$ measures the novelty of patent i. We assume $\mathbb{E}[\varepsilon_i | \mathbf{p}_{1,i} \cdot \mathbf{p}_{2,i}] = 0$ and $\text{Var}(\mathbf{p}_{1,i} \cdot \mathbf{p}_{2,i}) > 0$ so that OLS regression of Y_i on $\mathbf{p}_{1,i} \cdot \mathbf{p}_{2,i}$ would be consistent if $\mathbf{p}_{1,i}$ and $\mathbf{p}_{2,i}$ were observed.

As $\mathbf{p}_{1,i}$ and $\mathbf{p}_{2,i}$ are not observed, a pragmatic two-step strategy is to estimate γ_1 by regressing Y_i on $(\hat{\mathbf{p}}_{1,i} \cdot \hat{\mathbf{p}}_{2,i})$, where

$$\hat{\mathbf{p}}_{1,i} = \frac{\mathbf{x}_{1,i}}{C_{1,i}} \quad \text{and} \quad \hat{\mathbf{p}}_{2,i} = \frac{\mathbf{x}_{2,i}}{C_{2,i}}$$

are the term frequencies. Let $\hat{\gamma}_1$ denote the OLS estimator. To simplify derivations, we assume (i) Y_i, $\mathbf{x}_{1,i}$, and $\mathbf{x}_{2,i}$ are independent conditional on $(C_{1,i}, C_{2,i}, \mathbf{p}_{1,i}, \mathbf{p}_{2,i})$, and (ii) $C_{1,i}$ and $C_{2,i}$ are independent of each other and of $(\mathbf{p}_{1,i}, \mathbf{p}_{2,i}, Y_i)$. Consider a sequence of populations in which the conditional distribution of Y_i, $\mathbf{x}_{1,i}$, $\mathbf{x}_{2,i}$, $\mathbf{p}_{1,i}$, and $\mathbf{p}_{2,i}$ conditional on $(C_{1,i}, C_{2,i})$ is fixed, and the distribution of $(C_{1,i}, C_{2,i})$ grows with n so that

$$\sqrt{n}\,\mathbb{E}\left[\frac{1}{C_{t,i}}\right] \to \kappa_t \in [0, \infty) \tag{20}$$

for $t = 1, 2$.

Theorem 8. *Consider the sequence of populations just described. Then*

$$\sqrt{n}(\hat{\gamma}_1 - \gamma_1) \to_d N\left(\kappa_1 b_1 + \kappa_2 b_2, \frac{\mathbb{E}\left[\varepsilon_i^2(\mathbf{p}_{1,i} \cdot \mathbf{p}_{2,i} - \mathbb{E}\left[\mathbf{p}_{1,i} \cdot \mathbf{p}_{2,i}\right])^2\right]}{\mathrm{Var}(\mathbf{p}_{1,i} \cdot \mathbf{p}_{2,i})^2}\right),$$

where

$$b_1 = -\left(\frac{\mathbb{E}[\mathbf{p}_{2,i}^T(\mathrm{diag}(\mathbf{p}_{1,i}) - \mathbf{p}_{1,i}\mathbf{p}_{1,i}^T)\mathbf{p}_{2,i}]}{\mathrm{Var}(\mathbf{p}_{1,i} \cdot \mathbf{p}_{2,i})}\right)\gamma_1, \quad b_2 = -\left(\frac{\mathbb{E}[\mathbf{p}_{1,i}^T(\mathrm{diag}(\mathbf{p}_{2,i}) - \mathbf{p}_{2,i}\mathbf{p}_{2,i}^T)\mathbf{p}_{1,i}]}{\mathrm{Var}(\mathbf{p}_{1,i} \cdot \mathbf{p}_{2,i})}\right)\gamma_1.$$

As with our earlier results, Theorem 8 shows that $\hat{\mathbf{p}}_{1,i}$ and $\hat{\mathbf{p}}_{2,i}$ can be treated as if they are the true $\mathbf{p}_{1,i}$ and $\mathbf{p}_{2,i}$ for inference on γ_1 provided the amount of unstructured data per observation is "large" in the sense that both $\kappa_1 = 0$ and $\kappa_2 = 0$. For instance, in the patent example we would require both the amount of unstructured data per patent to be large (so that $\kappa_1 = 0$) and the amount of unstructured data for the existing stock to be large (so that $\kappa_2 = 0$). Otherwise, the location of two-step confidence intervals is shifted towards the origin, which leads to biased inference.

We focused on the simplest case of regression on dot-product similarity between term frequencies to simplify exposition. But our findings will extend to regression on other measures including cosine similarity, TF-IDF measures, transforms (e.g., logs) of similarity measures, and similarity measures formed from topic weights.

F Supplemental Results and Proofs

Notation Let $\|\cdot\|$ denote the Euclidean norm when applied to vectors and the spectral norm when applied to matrices. Let $\|\cdot\|_F$ denote the Frobenius norm.

F.1 Proofs for Section 2

Proof of Proposition 1. We start by writing

$$\begin{aligned}\sqrt{n}(\hat{\gamma}_1 - \gamma_1) &= \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n (Y_i - \gamma_1(\hat{\theta}_i - \bar{\hat{\theta}}))(\hat{\theta}_i - \bar{\hat{\theta}})}{\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \bar{\hat{\theta}})^2} \\ &= -\gamma_1 \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)(\hat{\theta}_i - \bar{\hat{\theta}})}{\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \bar{\hat{\theta}})^2} + \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i(\hat{\theta}_i - \bar{\hat{\theta}})}{\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \bar{\hat{\theta}})^2} \;=:\; T_{1,n} + T_{2,n},\end{aligned}$$

where $\bar{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^n \hat{\theta}_i$. Note by Chebyshev's inequality that for integers $k_1, k_2 \geq 0$ and any $t > 0$, we have

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n \hat{\theta}_i^{k_1}\theta_i^{k_2} - \mathbb{E}[\hat{\theta}_i^{k_1}\theta_i^{k_2}]\right| > t\right) \leq \frac{\mathbb{E}[\hat{\theta}_i^{2k_1}\theta_i^{2k_2}]}{t^2 n} \leq \frac{1}{t^2 n}. \tag{21}$$

Consider the denominator term in $T_{1,n}$ and $T_{2,n}$. By inequality (21), we have

$$\left|\frac{1}{n}\sum_{i=1}^n (\hat{\theta}_i - \bar{\hat{\theta}})^2 - \mathrm{Var}(\hat{\theta}_i)\right| \to_p 0,$$

where, by the law of total variance and independence of C_i and θ_i,

$$\mathrm{Var}(\hat{\theta}_i) = \mathrm{Var}(\theta_i) + \mathbb{E}\left[\frac{1}{C_i}\right]\mathbb{E}[\theta_i(1-\theta_i)] \to \mathrm{Var}(\theta_i)$$

because $\mathbb{E}\left[C_i^{-1}\right] \to 0$.

For the numerator in $T_{1,n}$, we similarly have

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)(\hat{\theta}_i - \bar{\hat{\theta}}) - \frac{1}{\sqrt{n}}\sum_{i=1}^n (\hat{\theta}_i - \theta_i)(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i])\right| \to_p 0.$$

Because $\mathbb{E}[\hat{\theta}_i|\theta_i] = \theta_i$ and $\mathrm{Var}(\hat{\theta}_i|\theta_i, C_i) = C_i^{-1}\theta_i(1-\theta_i)$, we have

$$\begin{aligned}\mathbb{E}\left[\sqrt{n}(\hat{\theta}_i - \theta_i)(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i])\right] &= \mathbb{E}\left[\sqrt{n}(\hat{\theta}_i - \theta_i)^2\right] \\ &= \sqrt{n}\,\mathbb{E}\left[\frac{1}{C_i}\right]\mathbb{E}[\theta_i(1-\theta_i)] \to \kappa\,\mathbb{E}[\theta_i(1-\theta_i)].\end{aligned}$$

A second application of inequality (21) gives

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\sqrt{n}(\hat{\theta}_i-\theta_i)(\hat{\theta}_i-\mathbb{E}[\hat{\theta}_i])-\mathbb{E}[\sqrt{n}(\hat{\theta}_i-\theta_i)(\hat{\theta}_i-\mathbb{E}[\hat{\theta}_i])]\right|>t\right)$$
$$\leq \frac{\mathbb{E}[(\hat{\theta}_i-\theta_i)^2(\hat{\theta}_i-\mathbb{E}[\hat{\theta}_i])^2]}{t^2}\leq\frac{\mathbb{E}[(\hat{\theta}_i-\theta_i)^2]}{t^2}=\mathbb{E}\left[\frac{1}{C_i}\right]\frac{\mathbb{E}[\theta_i(1-\theta_i)]}{t^2}\to 0\,.$$

Hence,

$$T_{1,n}\to_p -\kappa\,\gamma_1\frac{\mathbb{E}[\theta_i(1-\theta_i)]}{\mathrm{Var}(\theta_i)}\,.$$

For $T_{2,n}$, we have

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(\hat{\theta}_i-\bar{\hat{\theta}})-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(\hat{\theta}_i-\theta_i)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(\theta_i-\mathbb{E}[\theta_i])\right|\to_p 0$$

because $(\bar{\hat{\theta}}-\mathbb{E}[\theta_i])\times\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\to_p 0$. Note that $\mathbb{E}[\varepsilon_i(\hat{\theta}_i-\theta_i)]=0$ because Y_i and (X_i,C_i) are independent conditional on θ_i and both ε_i and $\hat{\theta}_i-\theta_i$ have conditional (on θ_i) mean zero. Hence by Chebyshev's inequality, for any $t>0$ we have

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}\sqrt{n}\varepsilon_i(\hat{\theta}_i-\theta_i)\right|>t\right)\leq\frac{\mathbb{E}[\varepsilon_i^2(\hat{\theta}_i-\theta_i)^2]}{t^2}=\mathbb{E}\left[\frac{1}{C_i}\right]\frac{\mathbb{E}[\varepsilon_i^2\theta_i(1-\theta_i)]}{t^2}\to 0,$$

because ε_i and (X_i,C_i) are independent conditional on θ_i, C_i and θ_i are independent, and $\mathbb{E}\left[C_i^{-1}\right]\to 0$. Finally, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(\theta_i-\mathbb{E}[\theta_i])$ is asymptotically $N(0,\mathbb{E}[\varepsilon_i^2(\theta_i-\mathbb{E}[\theta_i])^2])$ by the central limit theorem. ∎

F.2 Proofs for Appendix B

F.2.1 AI/ML-Generated Labels

Proof of Theorem 4. Assumption 1(i) holds by Assumption 3(ii) and the fact that $\|\boldsymbol{\theta}_i\|\leq 1$. The first part of Assumption 1(ii) holds by the law of large numbers and the fact that $\mathbb{E}[\|\boldsymbol{\xi}_i\|^2]<\infty$. For the second part, by Assumption 3(i) and the law of large numbers we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)^T&=\frac{1}{n}\sum_{i=1}^{n}(\mathbf{r}(\mathbf{x}_i,U_i)-\boldsymbol{\theta}_i)(\mathbf{r}(\mathbf{x}_i,U_i)-\boldsymbol{\theta}_i)^T+o_p(1)\\&\to_p\mathbb{E}\left[(\mathbf{r}(\mathbf{x}_i,U_i)-\boldsymbol{\theta}_i)(\mathbf{r}(\mathbf{x}_i,U_i)-\boldsymbol{\theta}_i)^T\right]\\&=\mathbb{E}\left[\mathrm{diag}(\boldsymbol{\pi}(\mathbf{x}_i)+\mathbf{p}(\mathbf{x}_i))-\boldsymbol{\pi}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^T-\mathbf{p}(\mathbf{x}_i)\boldsymbol{\pi}(\mathbf{x}_i)^T\right]=:\mathbf{V},\end{aligned}$$

where the final line is by independence of $\boldsymbol{\theta}_i$ and $\mathbf{r}(\mathbf{x}_i, U_i)$ conditional on $\mathbf{x}_i$. Similarly, for the third part, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\boldsymbol{\theta}_i^T &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i)\boldsymbol{\theta}_i^T + o_p(1) \\ &\to_p \mathbb{E}\left[(\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i)\boldsymbol{\theta}_i^T\right] \\ &= \mathbb{E}\left[\boldsymbol{\pi}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^T - \operatorname{diag}(\mathbf{p}(\mathbf{x}_i))\right] =: \mathbf{W}.\end{aligned}$$

For the fourth part, first note that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i - \mathbf{r}(\mathbf{x}_i, U_i))\mathbf{q}_i^T\right\| \leq \max_{1\leq i\leq n}\left\|\hat{\boldsymbol{\theta}}_i - \mathbf{r}(\mathbf{x}_i, U_i)\right\| \times \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{q}_i\| \to_p 0,$$

by Assumption 3(i)-(ii). Then by Assumption 3(ii) and the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i)\mathbf{q}_i^T \to_p \mathbb{E}\left[(\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i)\mathbf{q}_i^T\right],$$

which is zero by Assumption 3(iii).

For the final part of Assumption 1(ii), first note

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\varepsilon_i = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i\varepsilon_i \\ \frac{1}{n}\sum_{i=1}^{n}\mathbf{q}_i\varepsilon_i \end{bmatrix}, \tag{22}$$

where $\frac{1}{n}\sum_{i=1}^{n}\mathbf{q}_i\varepsilon_i \to_p \mathbf{0}$ by the law of large numbers and Assumption 3(ii). Moreover,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\theta}}_i - \mathbf{r}(\mathbf{x}_i, U_i)\right)\varepsilon_i\right\| \to_p 0,$$

by Assumption 3(i)-(ii). Finally,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{r}(\mathbf{x}_i, U_i)\varepsilon_i \to_p \mathbb{E}\left[\mathbf{r}(\mathbf{x}_i, U_i)\varepsilon_i\right] = \mathbf{0}$$

by the law of large numbers and independence of $\mathbf{r}(\mathbf{x}_i, U_i)$ and ε_i conditional on $\mathbf{x}_i$, and the fact that $\mathbb{E}[\varepsilon_i|\mathbf{x}_i] = 0$. ∎

The following lemma pertains to the sequence-of-populations asymptotic framework.

Lemma 3. *Let $\mathbf{z}_i$ be a random vector with finite fourth moment and let Assumption 4(i)*

and (17) hold. Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbf{z}_i(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)^T-\mathbb{E}\left[\mathbf{z}_i(\mathbf{r}(\mathbf{x}_i,U_i)-\boldsymbol{\theta}_i)^T\right]\right)\to_p 0.$$

Proof of Lemma 3. First note that

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{z}_i(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)^T-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{z}_i(\mathbf{r}(\mathbf{x}_i,U_i)-\boldsymbol{\theta}_i)^T\right\| \leq \sqrt{n}\max_{1\leq i\leq n}\|\hat{\boldsymbol{\theta}}_i-\mathbf{r}(\mathbf{x}_i,U_i)\|\times\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{z}_i\|\to_p 0,$$

by Assumption 4(i) and the fact that $\mathbb{E}[\|\mathbf{z}_i\|]<\infty$. Now let $\mathbf{X}_{i,n}=\mathbf{z}_i(\mathbf{r}(\mathbf{x}_i,U_i)-\boldsymbol{\theta}_i)^T-\mathbb{E}\left[\mathbf{z}_i(\mathbf{r}(\mathbf{x}_i,U_i)-\boldsymbol{\theta}_i)^T\right]$. With D denoting the dimension of $\mathbf{z}_i$, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_{i,n}\right\|_F^2\right] &= \sum_{j=1}^{D}\sum_{k=1}^{K}\mathbb{E}\left[(\mathbf{X}_{i,n})_{j,k}^2\right]\\ &\leq \sum_{j=1}^{D}\sum_{k=1}^{K}\mathbb{E}\left[(\mathbf{z}_{i,j})^2\left(r_k(\mathbf{x}_i,U_i)-\theta_{i,k}\right)^2\right]\\ &\leq \sum_{j=1}^{D}\sum_{k=1}^{K}\mathbb{E}\left[(\mathbf{z}_{i,j})^4\right]^{1/2}\mathbb{E}\left[\left(r_k(\mathbf{x}_i,U_i)-\theta_{i,k}\right)^4\right]^{1/2}\\ &\leq \text{constant}\times\sum_{k=1}^{K}\mathbb{E}\left[\left(r_k(\mathbf{x}_i,U_i)-\theta_{i,k}\right)^2\right]^{1/2}\to 0,\end{aligned}$$

where the second inequality is by Cauchy-Schwarz, the third is because $\mathbf{z}_i$ has finite fourth moment and the fact that $r_k(\mathbf{x}_i,U_i)-\theta_{i,k}$ takes values in $\{-1,0,1\}$, and convergence to zero is by (15) because $\mathbb{E}\left[\left(r_k(\mathbf{x}_i,U_i)-\theta_{i,k}\right)^2\right]=(\mathbf{V})_{k,k}$. The result now follows by Chebyshev's inequality. ∎

Proof of Theorem 5. Assumption 2(i) is implied by Assumption 4(ii).

We now verify Assumption 2(ii). First by Lemma 3 and Assumption 4, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)\mathbf{q}_i^T\to_p\mathbf{0}, \tag{23}$$

which establishes the final part of Assumption 2(ii). We may similarly deduce by Assumption 4(i) that

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)^T-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{r}(\mathbf{x}_i,U_i)(\mathbf{r}(\mathbf{x}_i,U_i)-\boldsymbol{\theta}_i)^T\right\|\to_p 0.$$

Further, with $\mathbf{X}_{i,n} = \mathbf{r}(\mathbf{x}_i, U_i)(\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i)^T - \mathbb{E}\left[\mathbf{r}(\mathbf{x}_i, U_i)(\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i)^T\right]$ we have by similar arguments to the proof of Lemma 3 that $\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{X}_{i,n} \to_p \mathbf{0}$. Hence it follows by (16) that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\boldsymbol{\theta}}_i(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T \to_p \kappa\,\boldsymbol{\Omega},$$

which establishes the second part of Assumption 2(ii). The preceding display also implies that

$$\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\theta}}_i(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T \to_p \mathbf{0},$$

and hence, by (23), that

$$\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\xi}}_i(\hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i)^T \to_p \mathbf{0}.$$

But note that

$$\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T = \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\xi}}_i(\hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i)^T + \frac{1}{n}\sum_{i=1}^n (\hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i)\boldsymbol{\xi}_i^T + \frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i\boldsymbol{\xi}_i^T.$$

We have just shown that the first term on the right-hand side is asymptotically negligible. Moreover, $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i\boldsymbol{\xi}_i^T \to_p \mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]$ by the law of large numbers and Assumption 4(ii). To verify the first part of Assumption 2(ii), it therefore remains to show that the second term on the right-hand side of the above display is asymptotically negligible. In view of (23) it is enough to show

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{\theta}_i(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T \to_p \mathbf{0}. \tag{24}$$

But by Lemma 3 using Assumption 4(i)-(ii), we have

$$\left\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{\theta}_i(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T - \mathbb{E}\left[\boldsymbol{\theta}_i(\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i)^T\right] \right\| \to_p 0.$$

Moreover, $\mathbb{E}\left[\boldsymbol{\theta}_i(\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i)^T\right] = \mathbb{E}\left[\mathbf{p}(\mathbf{x}_i)\boldsymbol{\pi}(\mathbf{x}_i)^T - \mathrm{diag}(\mathbf{p}(\mathbf{x}_i))\right] \to \mathbf{0}$ by (15) and the fact that $\mathbf{V}$ is diagonally dominant. This proves (24) and completes the verification of Assumption 2(ii).

Now consider Assumption 2(iii). For the first part, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (\hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i)\varepsilon_i = \begin{bmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^n (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\varepsilon_i \\ \mathbf{0} \end{bmatrix}.$$

By Lemma 3 using Assumption 4(i)-(ii) and $\mathbb{E}\left[(\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i)\varepsilon_i\right] = \mathbf{0}$, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\varepsilon_i \to_p \mathbf{0}.$$

It follows that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \varepsilon_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\xi}_i \varepsilon_i + o_p(1)$. The first part of Assumption 2(iii) now holds by the central limit theorem and Assumption 4(ii).

To complete the verification of Assumption 2(iii), we proceed as in the proof of Theorem 7. We start by writing

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \left(\hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T - \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T \right) + \frac{1}{n} \sum_{i=1}^{n} (\hat{\varepsilon}_i^2 - \varepsilon_i^2) \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T =: T_{1,n} + T_{2,n} + T_{3,n},$$

where again $T_{1,n} \to_p \mathbb{E}\left[\varepsilon_i^2 \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]$ by the LLN and Assumption 4(ii). To show $T_{2,n} \to_p \mathbf{0}$, it suffices to show

$$T_{2,a,n} := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \left(\hat{\boldsymbol{\theta}}_i \hat{\boldsymbol{\theta}}_i^T - \boldsymbol{\theta}_i \boldsymbol{\theta}_i^T \right) \to_p \mathbf{0}, \tag{25}$$

and

$$T_{2,b,n} := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \mathbf{q}_i \left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i \right)^T \to_p \mathbf{0}. \tag{26}$$

For $T_{2,a,n}$, we may first deduce by Assumption 4(i)-(ii) that

$$\left\| T_{2,a,n} - \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \left(\mathbf{r}(\mathbf{x}_i, U_i) \mathbf{r}(\mathbf{x}_i, U_i)^T - \boldsymbol{\theta}_i \boldsymbol{\theta}_i^T \right) \right\| \to_p 0.$$

Then since $\boldsymbol{\theta}_i \boldsymbol{\theta}_i^T = \operatorname{diag}(\boldsymbol{\theta}_i)$ and similarly for $\mathbf{r}(\mathbf{x}_i, U_i)$, we have by Cauchy–Schwarz that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \left(\mathbf{r}(\mathbf{x}_i, U_i) \mathbf{r}(\mathbf{x}_i, U_i)^T - \boldsymbol{\theta}_i \boldsymbol{\theta}_i^T \right) \right\| \leq \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^4 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \|\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i\|^2 \right)^{1/2},$$

where the first term on the right-hand side is $O_p(1)$ by Assumption 4(ii) and the second term is $o_p(1)$ by Markov's inequality because $\mathbb{E}\left[\|\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i\|^2\right] = \operatorname{tr}(\mathbf{V}) \to 0$ by (15), proving (25). For $T_{2,b,n}$, we may similarly deduce by Assumption 4(i)-(ii) that

$$\left\| T_{2,b,n} - \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \mathbf{q}_i \left(\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i \right)^T \right\| \to_p 0.$$

Then by Hölder's inequality, we have

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \mathbf{q}_i \left(\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i \right)^T \right\| \leq \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^4 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \|\mathbf{q}_i\|^4 \right)^{1/4} \left(\frac{1}{n} \sum_{i=1}^{n} \|\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i\|^4 \right)^{1/4}.$$

The first two terms on the right-hand side are $O_p(1)$ by Assumption 4(ii). For the third term, note that $\|\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i\|^4 \leq 2\|\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i\|^2$ because $\|\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i\|^2$ takes the value 0, 1, or 2. Hence, the third term is $o_p(1)$ by Markov's inequality because $\mathbb{E}\left[\|\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i\|^2\right] = \mathrm{tr}(\mathbf{V}) \to 0$ by (15), proving (26).

Finally, note $\hat{\varepsilon}_i^2 - \varepsilon_i^2 = (\boldsymbol{\xi}_i^T \boldsymbol{\psi} - \hat{\boldsymbol{\xi}}_i^T \hat{\boldsymbol{\psi}})^2 + 2\varepsilon_i(\boldsymbol{\xi}_i^T \boldsymbol{\psi} - \hat{\boldsymbol{\xi}}_i^T \hat{\boldsymbol{\psi}})$. Hence, to show $T_{3,n} \to_p \mathbf{0}$, it suffices to show

$$T_{3,a,n} := \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\xi}_i^T \boldsymbol{\psi} - \hat{\boldsymbol{\xi}}_i^T \hat{\boldsymbol{\psi}})^2 \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T \to_p \mathbf{0}, \tag{27}$$

and

$$T_{3,b,n} := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i (\boldsymbol{\xi}_i^T \boldsymbol{\psi} - \hat{\boldsymbol{\xi}}_i^T \hat{\boldsymbol{\psi}}) \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T \to_p \mathbf{0}. \tag{28}$$

By Hölder's inequality, we have

$$\|T_{3,a,n}\| \leq \left(\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\xi}_i^T \boldsymbol{\psi} - \hat{\boldsymbol{\xi}}_i^T \hat{\boldsymbol{\psi}})^4 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \|\hat{\boldsymbol{\xi}}_i\|^4 \right)^{1/2},$$

and

$$\|T_{4,a,n}\| \leq \left(\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\xi}_i^T \boldsymbol{\psi} - \hat{\boldsymbol{\xi}}_i^T \hat{\boldsymbol{\psi}})^4 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \|\hat{\boldsymbol{\xi}}_i\|^4 \right)^{1/4} \left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^4 \right)^{1/4},$$

where $\frac{1}{n} \sum_{i=1}^{n} \|\hat{\boldsymbol{\xi}}_i\|^4 = O_p(1)$ by Assumption 4(i)-(ii) and $\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^4 = O_p(1)$ by Assumption 4(ii). For the remaining term, notice that

$$\boldsymbol{\xi}_i^T \boldsymbol{\psi} - \hat{\boldsymbol{\xi}}_i^T \hat{\boldsymbol{\psi}} = (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T \boldsymbol{\gamma} + \hat{\boldsymbol{\xi}}_i^T (\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}).$$

Then, since $(a+b)^4 \leq 8a^4 + 8b^4$, we have

$$\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\xi}_i^T \boldsymbol{\psi} - \hat{\boldsymbol{\xi}}_i^T \hat{\boldsymbol{\psi}})^4 \leq \left(\frac{8}{n} \sum_{i=1}^{n} \|\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i\|^4 \right) \|\boldsymbol{\gamma}\|^4 + \left(\frac{8}{n} \sum_{i=1}^{n} \|\hat{\boldsymbol{\xi}}_i\|^4 \right) \|\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\|^4,$$

where the second term on the right-hand side is $o_p(1)$ by consistency of $\hat{\boldsymbol{\psi}}$ and the fact that $\frac{1}{n} \sum_{i=1}^{n} \|\hat{\boldsymbol{\xi}}_i\|^4 = O_p(1)$, as established above. For the first term on the right-hand side, we have $\frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i\|^4 = \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{\theta}_i - r(\mathbf{x}_i, U_i)\|^4 + o_p(1)$ by Assumption 4(i).

Then arguing as above we have $\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{\theta}_i - r(\mathbf{x}_i, U_i)\|^4 \to_p 0$, proving (27) and (28). ∎

F.2.2 Topic Models

The next two lemmas apply in both fixed-populations and sequences-of-populations.

Lemma 4. *Suppose that (5) holds. Then*

$$\mathbb{E}\left[\hat{\mathbf{p}}_i\hat{\mathbf{p}}_i^T \middle| C_i, \boldsymbol{w}_i\right] = \mathbf{B}^T\boldsymbol{w}_i\boldsymbol{w}_i^T\mathbf{B} + \frac{1}{C_i}\left(\operatorname{diag}(\mathbf{B}^T\boldsymbol{w}_i) - \mathbf{B}^T\boldsymbol{w}_i\boldsymbol{w}_i^T\mathbf{B}\right),$$

and

$$\mathbb{E}\left[(\hat{\mathbf{p}}_i - \mathbf{p}_i)(\hat{\mathbf{p}}_i - \mathbf{p}_i)^T \middle| C_i, \boldsymbol{w}_i\right] = \frac{1}{C_i}\left(\operatorname{diag}(\mathbf{B}^T\boldsymbol{w}_i) - \mathbf{B}^T\boldsymbol{w}_i\boldsymbol{w}_i^T\mathbf{B}\right).$$

Proof of Lemma 4. First note by (5) that

$$\begin{aligned}
\mathbb{E}\left[\hat{\mathbf{p}}_i\hat{\mathbf{p}}_i^T \middle| C_i, \boldsymbol{w}_i\right] &= \frac{1}{C_i^2}\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^T \middle| C_i, \boldsymbol{w}_i\right] \\
&= \frac{1}{C_i^2}\left(\mathbb{E}\left[\mathbf{x}_i \middle| C_i, \boldsymbol{w}_i\right]\mathbb{E}\left[\mathbf{x}_i \middle| C_i, \boldsymbol{w}_i\right]^T + \operatorname{Var}\left[\mathbf{x}_i \middle| C_i, \boldsymbol{w}_i\right]\right) \\
&= \mathbf{B}^T\boldsymbol{w}_i\boldsymbol{w}_i^T\mathbf{B} + \frac{1}{C_i}\left(\operatorname{diag}(\mathbf{B}^T\boldsymbol{w}_i) - \mathbf{B}^T\boldsymbol{w}_i\boldsymbol{w}_i^T\mathbf{B}\right),
\end{aligned}$$

where the last line follows from the mean and variance of the multinomial distribution. The second result now follows because $\mathbb{E}\left[\hat{\mathbf{p}}_i \middle| C_i, \boldsymbol{w}_i\right] = \mathbf{p}_i = \mathbf{B}^T\boldsymbol{w}_i$. ∎

Lemma 5. *Let Assumption 5(i)-(iii) hold. Then*

$$\begin{aligned}
\Bigg\| &\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i\hat{\boldsymbol{\theta}}_i^T - \mathbb{E}\left[\boldsymbol{\theta}_i\boldsymbol{\theta}_i^T\right] \\
&- \mathbb{E}\left[\frac{1}{C_i}\right]\left(\boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\operatorname{diag}(\mathbf{B}^T\mathbb{E}[\boldsymbol{w}_i])\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\boldsymbol{S}^T - \mathbb{E}\left[\boldsymbol{\theta}_i\boldsymbol{\theta}_i^T\right]\right)\Bigg\| \to_p 0.
\end{aligned}$$

Proof of Lemma 5. In view of Assumption 5(iii), we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i\hat{\boldsymbol{\theta}}_i^T - \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{p}}_i\hat{\mathbf{p}}_i^T\right)\hat{\mathbf{B}}^T(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\boldsymbol{S}^T\right\| \to_p 0$$

where $(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}$ exists with probability approaching one by Assumption 5(i)-(ii). Each element of $\hat{\mathbf{p}}_i\hat{\mathbf{p}}_i^T$ is bounded between 0 and 1, so we may deduce by Chebyshev's inequality that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{p}}_i\hat{\mathbf{p}}_i^T - \mathbb{E}\left[\hat{\mathbf{p}}_i\hat{\mathbf{p}}_i^T\right]\right\| \to_p 0.$$

Hence, by Assumption 5(ii) and Slutsky's theorem, we have

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\theta}}_i \hat{\boldsymbol{\theta}}_i^T - \boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\,\mathbb{E}\left[\hat{\mathbf{p}}_i \hat{\mathbf{p}}_i^T\right]\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\boldsymbol{S}^T \right\| \to_p 0.$$

The result follows by Lemma 4 and independence of C_i and $\boldsymbol{\theta}_i$. ∎

Proof of Theorem 6. Assumption 1(i) holds by Assumption 5(iv) and the fact that $\boldsymbol{\theta}_i = \boldsymbol{S}\boldsymbol{w}_i$ where $\boldsymbol{w}_i$ takes values in Δ^{K-1}, hence $\|\boldsymbol{\theta}_i\| \leq 1$.

The first part of Assumption 1(ii) holds by the law of large numbers and the fact that $\mathbb{E}[\|\boldsymbol{\xi}_i\|^2] < \infty$. For the second part, we have

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\theta}}_i \hat{\boldsymbol{\theta}}_i^T \to_p \mathbb{E}\left[\boldsymbol{\theta}_i \boldsymbol{\theta}_i^T\right] + \mathbf{V}$$

by Lemma 5, which provides the expression for $\mathbf{V}$. Further, $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\theta}_i \boldsymbol{\theta}_i^T \to_p \mathbb{E}\left[\boldsymbol{\theta}_i \boldsymbol{\theta}_i^T\right]$ by the law of large numbers. To complete the proof of the second, third, and fourth parts of Assumption 1(ii), it suffices to show that

$$\frac{1}{n}\sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\boldsymbol{\xi}_i^T \to_p \mathbf{0}. \tag{29}$$

To this end, in view of Assumption 5(iii)-(iv), we have

$$\begin{aligned}\left\| \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\theta}}_i \boldsymbol{\xi}_i^T - \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{p}}_i \boldsymbol{\xi}_i^T\right) \right\| \\ \leq \left(\max_{1\leq i\leq n} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\hat{\mathbf{p}}_i\| \right) \times \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{\xi}_i\| \to_p 0.\end{aligned}$$

Note $(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}} \to_p (\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}$ by Assumption 5(i)-(ii), and $\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{p}}_i \boldsymbol{\xi}_i^T = O_p(1)$ by Assumption 5(iv) and the fact that $\|\hat{\mathbf{p}}_i \boldsymbol{\xi}_i^T\| \leq \|\boldsymbol{\xi}_i\|$. Hence,

$$\begin{aligned}\left\| \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{p}}_i \boldsymbol{\xi}_i^T\right) - \boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{p}}_i \boldsymbol{\xi}_i^T\right) \right\| \\ \leq \|\boldsymbol{S}\| \left\| (\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}} - (\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B} \right\| \times \left\| \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{p}}_i \boldsymbol{\xi}_i^T \right\| \to_p 0.\end{aligned}$$

Finally,

$$\left\| \boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{p}}_i \boldsymbol{\xi}_i^T\right) - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\theta}_i \boldsymbol{\xi}_i^T \right\| = \left\| \boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B} \left(\frac{1}{n}\sum_{i=1}^{n} (\hat{\mathbf{p}}_i - \mathbf{p_i})\boldsymbol{\xi}_i^T\right) \right\|.$$

But

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{p}}_i - \mathbf{p_i})\boldsymbol{\xi}_i^T \to_p \mathbb{E}\left[(\hat{\mathbf{p}}_i - \mathbf{p_i})\boldsymbol{\xi}_i^T\right] = \mathbf{0}$$

by the law of large numbers, independence of $\mathbf{x}_i$ and $\mathbf{q}_i$ conditional on $(C_i, \boldsymbol{w}_i)$, and the fact that $\mathbb{E}[\hat{\mathbf{p}}_i|C_i, \boldsymbol{w}_i] = \mathbf{p}_i$. This proves (29), from which we also conclude that $\mathbf{W} = \mathbf{0}$.

To verify the final part of Assumption 1(ii), first note

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\varepsilon_i = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i\varepsilon_i \\ \frac{1}{n}\sum_{i=1}^{n}\mathbf{q}_i\varepsilon_i \end{bmatrix}, \tag{30}$$

where $\frac{1}{n}\sum_{i=1}^{n}\mathbf{q}_i\varepsilon_i \to_p \mathbf{0}$ by the law of large numbers and Assumption 5(iv). For the remaining term, we use Assumption 5(iii)-(iv) to deduce

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i\varepsilon_i - \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{p}}_i\varepsilon_i\right)\right\| \to_p 0.$$

We have $(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}} \to_p (\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}$ by Assumption 5(i)-(ii). Moreover, Assumption 5(iv) and the fact that $\hat{\mathbf{p}}_i$ takes values in the simplex imply that $\mathbb{E}[\|\hat{\mathbf{p}}_i\varepsilon_i\|] < \infty$. Hence,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{p}}_i\varepsilon_i \to_p \mathbb{E}\left[\hat{\mathbf{p}}_i\varepsilon_i\right] = \mathbf{0}$$

by the law of large numbers, independence of $(\mathbf{x}_i, C_i)$ and ε_i conditional on $(\boldsymbol{w}_i, \mathbf{q}_i)$, and the fact that $\mathbb{E}[\varepsilon_i|\boldsymbol{w}_i, \mathbf{q}_i] = 0$. ∎

The next three lemmas are used in the proof of Theorem 7. They are derived in the sequence-of-populations asymptotic framework in which (18) holds.

Lemma 6. *Let Assumption 6(i)-(iii) hold. Then*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T \to_p -\kappa\left(\boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\,\mathrm{diag}(\mathbf{B}^T\mathbb{E}[\boldsymbol{w}_i])\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\boldsymbol{S}^T - \mathbb{E}[\boldsymbol{\theta}_i\boldsymbol{\theta}_i^T]\right).$$

Proof of Lemma 6. First note that $\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T - T_{1,n} - T_{2,n}\right\| \to_p 0$ by Assumption 6(iii), where

$$T_{1,n} = \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\left(\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{p}}_i\mathbf{p}_i^T\right)\sqrt{n}\left(\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1} - \hat{\mathbf{B}}^T(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\right)\right)\boldsymbol{S}^T$$

$$T_{2,n} = \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\mathbf{p}}_i(\mathbf{p}_i - \hat{\mathbf{p}}_i)^T\right)\hat{\mathbf{B}}^T(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\boldsymbol{S}^T.$$

Assumption 6(i)-(ii) implies that $\sqrt{n}(\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1} - \hat{\mathbf{B}}^T(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}) \to_p \mathbf{0}$. Moreover, as $\|\frac{1}{n}\sum_{i=1}^n \hat{\mathbf{p}}_i \mathbf{p}_i^T\| \leq 1$, it follows that $T_{1,n} \to_p \mathbf{0}$.

For term $T_{2,n}$, note by Lemma 4 that

$$\begin{aligned}\mathbb{E}\left[\hat{\mathbf{p}}_i\left(\hat{\mathbf{p}}_i - \mathbf{p}_i\right)^T\right] &= \mathbb{E}\left[\left(\hat{\mathbf{p}}_i - \mathbf{p}_i\right)\left(\hat{\mathbf{p}}_i - \mathbf{p}_i\right)^T\right] \\ &= \mathbb{E}\left[\frac{1}{C_i}\right]\left(\operatorname{diag}\left(\mathbf{B}^T\mathbb{E}\left[\boldsymbol{w}_i\right]\right) - \mathbf{B}^T\mathbb{E}\left[\boldsymbol{w}_i\boldsymbol{w}_i^T\right]\mathbf{B}\right). \end{aligned} \tag{31}$$

Let $\mathbf{X}_i = \hat{\mathbf{p}}_i\left(\hat{\mathbf{p}}_i - \mathbf{p}_i\right)^T - \mathbb{E}\left[\hat{\mathbf{p}}_i\left(\hat{\mathbf{p}}_i - \mathbf{p}_i\right)^T\right]$. Then with $\|\cdot\|_F$ denoting the Frobenius norm,

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{X}_i\right\|_F^2\right] = \sum_{j=1}^V\sum_{k=1}^V \mathbb{E}\left[\left(\mathbf{X}_i\right)_{j,k}^2\right] &\leq \sum_{j=1}^V\sum_{k=1}^V \mathbb{E}\left[\left(\hat{\mathbf{p}}_{i,j}\right)^2\left(\hat{\mathbf{p}}_{i,k} - \mathbf{p}_{i,k}\right)^2\right] \\ &\leq \sum_{k=1}^V \mathbb{E}\left[\left(\hat{\mathbf{p}}_{i,k} - \mathbf{p}_{i,k}\right)^2\right] \to 0,\end{aligned}$$

where the second inequality is because $\hat{\mathbf{p}}_i$ is in the simplex and the convergence to zero holds in view of (18) and (31). It follows that

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{\mathbf{p}}_i(\mathbf{p}_i - \hat{\mathbf{p}}_i)^T - \sqrt{n}\,\mathbb{E}\left[\hat{\mathbf{p}}_i\left(\mathbf{p}_i - \hat{\mathbf{p}}_i\right)^T\right]\right\| \to_p 0.$$

We conclude that $T_{2,n} \to_p -\kappa\left(\boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\operatorname{diag}(\mathbf{B}^T\mathbb{E}[\boldsymbol{w}_i])\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\boldsymbol{S}^T - \mathbb{E}[\boldsymbol{w}_i\boldsymbol{w}_i^T]\right)$ by (18), (31), and Assumption 6(i)-(ii) ∎

Lemma 7. *Let Assumption 6(i)-(iv) hold. Then*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\xi}_i(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T \to_p \mathbf{0}.$$

Proof of Lemma 7. First note that $\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\xi}_i(\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T - T_{1,n} - T_{2,n}\right\| \to_p 0$ by Assumption 6(iii)-(iv), where

$$\begin{aligned}T_{1,n} &= \left(\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i\mathbf{p}_i^T\right)\sqrt{n}\left(\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1} - \hat{\mathbf{B}}^T(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\right)\right)\boldsymbol{S}^T \\ T_{2,n} &= \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\xi}_i(\mathbf{p}_i - \hat{\mathbf{p}}_i)^T\right)\hat{\mathbf{B}}^T(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\boldsymbol{S}^T.\end{aligned}$$

Assumption 6(i)-(ii) implies that $\sqrt{n}(\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1} - \hat{\mathbf{B}}^T(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}) \to_p \mathbf{0}$. Moreover, $\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i\,\mathbf{p}_i^T\| \leq \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{\xi}_i\|$, which is bounded in probability by Assumption 6(iv). It follows that $T_{1,n} \to_p \mathbf{0}$. For $T_{2,n}$, note that $\mathbb{E}[\boldsymbol{\xi}_i(\hat{\mathbf{p}}_i - \mathbf{p}_i)^T] = \mathbf{0}$ by independence of $\mathbf{x}_i$

and $\mathbf{q}_i$ conditional on $(C_i, \boldsymbol{w}_i)$ and the fact that $\mathbb{E}[\hat{\mathbf{p}}_i | C_i, \boldsymbol{w}_i] = \mathbf{p}_i$. Let $\mathbf{X}_i = \boldsymbol{\xi}_i \left(\hat{\mathbf{p}}_i - \mathbf{p}_i\right)^T$ and let D denote the dimension of $\boldsymbol{\xi}_i$. Then

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i\right\|_F^2\right] = \sum_{j=1}^{D}\sum_{k=1}^{V}\mathbb{E}\left[\left(\mathbf{X}_i\right)_{j,k}^2\right] &= \sum_{j=1}^{D}\sum_{k=1}^{V}\mathbb{E}\left[\left(\boldsymbol{\xi}_{i,j}\right)^2\left(\hat{\mathbf{p}}_{i,k} - \mathbf{p}_{i,k}\right)^2\right] \\
&\leq \sum_{j=1}^{D}\sum_{k=1}^{V}\mathbb{E}\left[\left(\boldsymbol{\xi}_{i,j}\right)^4\right]^{1/2}\mathbb{E}\left[\left(\hat{\mathbf{p}}_{i,k} - \mathbf{p}_{i,k}\right)^4\right]^{1/2} \\
&\leq \text{constant} \times \sum_{k=1}^{V}\mathbb{E}\left[\left(\hat{\mathbf{p}}_{i,k} - \mathbf{p}_{i,k}\right)^2\right]^{1/2} \to 0,
\end{aligned}$$

where the first inequality is by Cauchy-Schwarz, the second is by Assumption 6(iv) and the fact that $|\hat{\mathbf{p}}_{i,k} - \mathbf{p}_{i,k}| \leq 1$, and convergence to zero is by (18) and (31). It follows that $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i \to_p \mathbf{0}$. We conclude by Assumption 6(i)-(ii) that $T_{2,n} \to_p \mathbf{0}$. ∎

Lemma 8. *Let Assumption 6(v) hold. Then*

$$\max_{1\leq i\leq n}\|\hat{\mathbf{p}}_i - \mathbf{p}_i\| \to_p 0.$$

Proof of Lemma 8. Let $\|\cdot\|_1$ be the ℓ^1 norm. As $\hat{\mathbf{p}}_i | (C_i, \boldsymbol{w}_i) \sim {C_i}^{-1}\text{Multinomial}(C_i, \mathbf{p}_i)$, for all $t > 0$ we have

$$\Pr\left(\max_{1\leq i\leq n}\|\hat{\mathbf{p}}_i - \mathbf{p}_i\|_1 > t \,\middle|\, \{(C_i, \boldsymbol{w}_i)\}_{i=1}^n\right) \leq \sum_{i=1}^{n}(2^V - 2)\exp\left\{-\frac{C_i t^2}{2K}\right\}$$

by the union bound and Lemma 1 of Mardia et al. (2019). Then by Assumption 6(v),

$$\Pr\left(\max_{1\leq i\leq n}\|\hat{\mathbf{p}}_i - \mathbf{p}_i\|_1 > t\right) \leq n(2^V - 2)\exp\left\{-\frac{c(\log n)^{1+\epsilon}t^2}{2K}\right\},$$

where $c, \epsilon > 0$. Hence, $\max_{1\leq i\leq n}\|\hat{\mathbf{p}}_i - \mathbf{p}_i\|_1 \to_p 0$. The result now follows because the ℓ^1 norm is weakly greater than the Euclidean norm. ∎

Proof of Theorem 7. Assumption 2(i) is implied by Assumption 6(iv).

We now verify Assumption 2(ii). The second and third parts of Assumption 2(ii) hold by Lemmas 6 and 7. For the first part, we have

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i(\hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i)^T + \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i)\boldsymbol{\xi}_i^T + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T.$$

The third term converges in probability to $\mathbb{E}\left[\boldsymbol{\xi}_i\,\boldsymbol{\xi}_i^T\right]$ by the law of large numbers and Assumption 6(iv), while the first two terms are $o_p(1)$ by Lemmas 6 and 7.

Now consider Assumption 2(iii). For the first part, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\varepsilon_i = \begin{bmatrix} \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i\varepsilon_i \\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{q}_i\varepsilon_i \end{bmatrix}.$$

We may deduce by arguments similar to those in the proof of Lemma 6 that

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i\varepsilon_i - \boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\mathbf{p}}_i\varepsilon_i\right)\right\| \to_p 0.$$

under Assumption 6(i)-(iii). Moreover,

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_i^2\left\|\hat{\mathbf{p}}_i-\mathbf{p}_i\right\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\varepsilon_i^2\middle|\boldsymbol{w}_i,\mathbf{q}_i\right]\mathbb{E}\left[\left\|\hat{\mathbf{p}}_i-\mathbf{p}_i\right\|^2\middle|\boldsymbol{w}_i,\mathbf{q}_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\varepsilon_i^2\middle|\boldsymbol{w}_i,\mathbf{q}_i\right]\mathbb{E}\left[\mathbb{E}\left[\left\|\hat{\mathbf{p}}_i-\mathbf{p}_i\right\|^2\middle|\boldsymbol{w}_i,C_i\right]\middle|\boldsymbol{w}_i,\mathbf{q}_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\varepsilon_i^2\middle|\boldsymbol{w}_i,\mathbf{q}_i\right]\mathbb{E}\left[\frac{1}{C_i}\mathrm{tr}\left\{\mathrm{diag}(\mathbf{B}^T\boldsymbol{w}_i)-\mathbf{B}^T\boldsymbol{w}_i\boldsymbol{w}_i^T\mathbf{B}\right\}\middle|\boldsymbol{w}_i,\mathbf{q}_i\right]\right] \\
&= \mathbb{E}\left[\frac{1}{C_i}\right]\mathbb{E}\left[\varepsilon_i^2\left(\mathrm{diag}(\mathbf{B}^T\boldsymbol{w}_i)-\mathbf{B}^T\boldsymbol{w}_i\boldsymbol{w}_i^T\mathbf{B}\right)\right] \to 0.
\end{aligned}$$

In the above display, the first equality is by independence of $(\mathbf{x}_i, C_i)$ and ε_i conditional on $(\boldsymbol{w}_i, \mathbf{q}_i)$, the second is by iterated expectations and independence of $\mathbf{x}_i$ and $\mathbf{q}_i$ conditional on $(C_i, \boldsymbol{w}_i)$, the third is by Lemma 4, and the fourth is by independence of C_i and $(Y_i, \mathbf{q}_i, \boldsymbol{w}_i)$ and condition (18). Therefore,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\mathbf{p}}_i-\mathbf{p}_i)\varepsilon_i \to_p 0$$

and so

$$\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\varepsilon_i - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\xi}_i\varepsilon_i\right\| \to_p 0.$$

It follows by the central limit theorem and Assumption 6(iv) that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\varepsilon_i \to_d N\left(\mathbf{0}, \mathbb{E}\left[\varepsilon_i^2\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]\right).$$

To complete verification of Assumption 2(iii), it remains to show

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T \to_p \mathbb{E}\left[\varepsilon_i^2\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right].$$

To this end, first write

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\left(\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T - \boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right) + \frac{1}{n}\sum_{i=1}^{n}(\hat{\varepsilon}_i^2 - \varepsilon_i^2)\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T =: T_{1,n} + T_{2,n} + T_{3,n}.$$

Evidently, $T_{1,n} \to_p \mathbb{E}\left[\varepsilon_i^2\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]$ by the LLN and Assumption 6(iv). For $T_{2,n}$, we have

$$T_{2,n} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2(\hat{\boldsymbol{\theta}}_i\hat{\boldsymbol{\theta}}_i^T - \boldsymbol{\theta}_i\boldsymbol{\theta}_i^T) & \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\mathbf{q}_i^T \\ \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\mathbf{q}_i(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T & \mathbf{0} \end{bmatrix}.$$

Consider the upper-left block. We may deduce by arguments similar to those in the proof of Lemma 6 that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2(\hat{\boldsymbol{\theta}}_i\hat{\boldsymbol{\theta}}_i^T - \boldsymbol{\theta}_i\boldsymbol{\theta}_i^T) - \boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2(\hat{\mathbf{p}}_i\hat{\mathbf{p}}_i^T - \mathbf{p}_i\mathbf{p}_i^T)\right)\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\boldsymbol{S}^T\right\| \to_p 0,$$

by Assumption 6(i)-(iv). Since $\mathbf{p}_i$ and $\hat{\mathbf{p}}_i$ both take values in the simplex, we have $\|\hat{\mathbf{p}}_i\hat{\mathbf{p}}_i^T - \mathbf{p}_i\mathbf{p}_i^T\| \leq 2\|\hat{\mathbf{p}}_i - \mathbf{p}_i\|$ and so

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2(\hat{\mathbf{p}}_i\hat{\mathbf{p}}_i^T - \mathbf{p}_i\mathbf{p}_i^T)\right\| \leq 2\left(\max_{1\leq i\leq n}\|\hat{\mathbf{p}}_i - \mathbf{p}_i\|\right)\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 \to_p 0,$$

by Lemma 8 and Assumption 6(iv). Now consider the off-diagonal blocks. By arguments similar to those in the proof of Lemma 7, we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\mathbf{q}_i(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T - \left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\mathbf{q}_i(\hat{\mathbf{p}}_i - \mathbf{p}_i)^T\right)\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\boldsymbol{S}^T\right\| \to_p 0,$$

by Assumption 6(i)-(iv). But note that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\mathbf{q}_i(\hat{\mathbf{p}}_i - \mathbf{p}_i)^T\right\| \leq \left(\max_{1\leq i\leq n}\|\hat{\mathbf{p}}_i - \mathbf{p}_i\|\right)\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\|\mathbf{q}_i\| \to_p 0,$$

by Lemma 8 and Assumption 6(iv). Therefore, $T_{2,n} \to_p \mathbf{0}$.

Now consider $T_{3,n}$. We have

$$\hat{\varepsilon}_i - \varepsilon_i = \hat{\boldsymbol{\theta}}_i^T(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}) + (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T\boldsymbol{\gamma} + \mathbf{q}_i^T(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}),$$

where

$$\max_{1\leq i\leq n}\left|\hat{\boldsymbol{\theta}}_i^T(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\right|$$
$$\leq\left(\max_{1\leq i\leq n}\|(\hat{\boldsymbol{\theta}}_i-\boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\hat{\mathbf{p}}_i)\|+\|\boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\|\right)\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|\to_p 0,$$

by Assumption 6(iii), consistency of $\hat{\boldsymbol{\gamma}}$, the fact that $\|\hat{\mathbf{p}}_i\|\leq 1$, and that $\|(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\|$ is bounded in probability by Assumption 6(i)-(ii). Moreover,

$$\max_{1\leq i\leq n}|(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)^T\boldsymbol{\gamma}|\leq\left(\max_{1\leq i\leq n}\|\hat{\boldsymbol{\theta}}_i-\boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\hat{\mathbf{p}}_i\|\right.$$
$$\left.+\|\boldsymbol{S}\|\left\|(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}-(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\right\|+\left\|\boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\right\|\max_{1\leq i\leq n}\|\hat{\mathbf{p}}_i-\mathbf{p}_i\|\right)\|\boldsymbol{\gamma}\|.$$

Consider the three terms in parentheses on the right-hand side of this display. The first two terms converge in probability to zero by Assumption 6(i)-(iii) and the third converges in probability to zero by Lemma 8. Finally, we have

$$\max_{1\leq i\leq n}\left\|\mathbf{q}_i^T(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}})\right\|\leq\left(\max_{1\leq i\leq n}\|\mathbf{q}_i\|\right)\|\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\|\to_p 0$$

by $\sqrt{n}$-consistency of $\hat{\boldsymbol{\alpha}}$ and the fact that $n^{-1/4}\max_{1\leq i\leq n}\|\mathbf{q}_i\|\to_p 0$ by Assumption 6(iv). Therefore, $\max_{1\leq i\leq n}|\hat{\varepsilon}_i-\varepsilon_i|\to_p 0$.

Now, since

$$\hat{\varepsilon}_i^2-\varepsilon_i^2=2(\hat{\varepsilon}_i-\varepsilon_i)\varepsilon_i+(\hat{\varepsilon}_i-\varepsilon_i)^2,$$

we have

$$T_{3,n}=\frac{2}{n}\sum_{i=1}^n(\hat{\varepsilon}_i-\varepsilon_i)\varepsilon_i\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T+\frac{1}{n}\sum_{i=1}^n(\hat{\varepsilon}_i-\varepsilon_i)^2\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T,$$

and so

$$\|T_{3,n}\|\leq 2\left(\max_{1\leq i\leq n}|\hat{\varepsilon}_i-\varepsilon_i|\right)\frac{1}{n}\sum_{i=1}^n|\varepsilon_i|\|\hat{\boldsymbol{\xi}}_i\|^2+\left(\max_{1\leq i\leq n}|\hat{\varepsilon}_i-\varepsilon_i|^2\right)\frac{1}{n}\sum_{i=1}^n\|\hat{\boldsymbol{\xi}}_i\|^2$$
$$=\left(\max_{1\leq i\leq n}|\hat{\varepsilon}_i-\varepsilon_i|\right)\operatorname{tr}\left\{\frac{2}{n}\sum_{i=1}^n|\varepsilon_i|\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right\}+\left(\max_{1\leq i\leq n}|\hat{\varepsilon}_i-\varepsilon_i|^2\right)\operatorname{tr}\left\{\frac{1}{n}\sum_{i=1}^n\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right\}\to_p 0,$$

because $\frac{1}{n}\sum_{i=1}^n|\varepsilon_i|\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T$ is bounded in probability by control of $T_{1,n}$ and $T_{2,n}$, which together imply $\frac{1}{n}\sum_{i=1}^n\varepsilon_i^2\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T$ is bounded in probability, and $\frac{1}{n}\sum_{i=1}^n\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T$ is bounded in probability by Lemmas 6 and 7, and Assumption 6(iv). ∎

F.3 Proofs for Appendix E

Proof of Theorem 8. Let $S_i = \mathbf{p}_{1,i} \cdot \mathbf{p}_{2,i}$ and $\hat{S}_i = \hat{\mathbf{p}}_{1,i} \cdot \hat{\mathbf{p}}_{2,i}$. By standard OLS algebra,

$$\sqrt{n}(\hat{\gamma}_1 - \gamma_1) = \frac{1}{\frac{1}{n}\sum_{i=1}^n (\hat{S}_i - \bar{\hat{S}})^2} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (\hat{S}_i - \bar{\hat{S}}) - \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (\hat{S}_i - S_i)(\hat{S}_i - \bar{\hat{S}}) \right) \gamma_1 \right),$$

where $\bar{\hat{S}} = \frac{1}{n}\sum_{i=1}^n \hat{S}_i$.

By Chebyshev's inequality, for all integers $k_1, k_2 \geq 0$ and all $t > 0$, we have

$$\Pr\left(\left| \frac{1}{n} \sum_{i=1}^n \hat{S}_i^{k_1} S_i^{k_2} - \mathbb{E}\left[\hat{S}_i^{k_1} S_i^{k_2} \right] \right| > t \right) \leq \frac{1}{nt^2} \mathbb{E}\left[\hat{S}_i^{2k_1} S_i^{2k_2} \right] \leq \frac{1}{nt^2}, \tag{32}$$

because $|\hat{S}_i| \leq \|\hat{\mathbf{p}}_{1,i}\| \|\hat{\mathbf{p}}_{2,i}\| \leq 1$ by virtue of the fact that $\|\hat{\mathbf{p}}_{t,i}\| \leq \|\hat{\mathbf{p}}_{t,i}\|_1 = 1$ for $t = 1, 2$, with $\|\cdot\|_1$ denoting the ℓ^1 norm, and similarly for S_i. Let $F_i = (\mathbf{p}_{1,i}, \mathbf{p}_{2,i}, C_{1,i}, C_{2,i})$. Note that

$$\mathbb{E}\left[\hat{S}_i \middle| F_i \right] = S_i$$

because $\mathbf{x}_{1,i}$, and $\mathbf{x}_{2,i}$ are independent conditional on $(C_{1,i}, C_{2,i}, \mathbf{p}_{1,i}, \mathbf{p}_{2,i})$. Hence, $\mathbb{E}[\hat{S}_i] = \mathbb{E}[S_i]$ and so it follows by (32) that $\bar{\hat{S}} \to_p \mathbb{E}[S_i]$. Moreover, for any conformable non-stochastic matrix $\mathbf{M}$, we have for $t = 1, 2$ that

$$\mathbb{E}\left[\hat{\mathbf{p}}_{t,i}^T \mathbf{M} \hat{\mathbf{p}}_{t,i} \middle| F_i \right] = \mathbf{p}_{t,i}^T \mathbf{M} \mathbf{p}_{t,i} + \frac{1}{C_{t,i}} \operatorname{tr}\left\{ \mathbf{M} \left(\operatorname{diag}(\mathbf{p}_{t,i}) - \mathbf{p}_{t,i}\mathbf{p}_{t,i}^T \right) \right\}.$$

Hence, by independence of $\mathbf{x}_{1,i}$ and $\mathbf{x}_{2,i}$ conditional on F_i, we have

$$\begin{aligned}
\mathbb{E}\left[\hat{S}_i^2 \middle| F_i \right] &= \mathbb{E}\left[\mathbb{E}\left[\hat{\mathbf{p}}_{1,i}^T (\hat{\mathbf{p}}_{2,i}\hat{\mathbf{p}}_{2,i}^T) \hat{\mathbf{p}}_{1,i} \middle| \hat{\mathbf{p}}_{2,i}, F_i \right] \middle| F_i \right] \\
&= \mathbb{E}\left[\mathbf{p}_{1,i}^T (\hat{\mathbf{p}}_{2,i}\hat{\mathbf{p}}_{2,i}^T) \mathbf{p}_{1,i} + \frac{1}{C_{1,i}} \operatorname{tr}\left\{ (\hat{\mathbf{p}}_{2,i}\hat{\mathbf{p}}_{2,i}^T) \left(\operatorname{diag}(\mathbf{p}_{1,i}) - \mathbf{p}_{1,i}\mathbf{p}_{1,i}^T \right) \right\} \middle| F_i \right] \\
&= \mathbb{E}\left[\hat{\mathbf{p}}_{2,i}^T (\mathbf{p}_{1,i}\mathbf{p}_{1,i}^T) \hat{\mathbf{p}}_{2,i} \middle| F_i \right] + \frac{1}{C_{1,i}} \mathbb{E}\left[\hat{\mathbf{p}}_{2,i}^T \left(\operatorname{diag}(\mathbf{p}_{1,i}) - \mathbf{p}_{1,i}\mathbf{p}_{1,i}^T \right) \hat{\mathbf{p}}_{2,i} \middle| F_i \right] \\
&= S_i^2 + \frac{1}{C_{2,i}} \mathbf{p}_{1,i}^T \left(\operatorname{diag}(\mathbf{p}_{2,i}) - \mathbf{p}_{2,i}\mathbf{p}_{2,i}^T \right) \mathbf{p}_{1,i} + \frac{1}{C_{1,i}} \mathbf{p}_{2,i}^T \left(\operatorname{diag}(\mathbf{p}_{1,i}) - \mathbf{p}_{1,i}\mathbf{p}_{1,i}^T \right) \mathbf{p}_{2,i} \\
&\quad + \frac{1}{C_{1,i}C_{2,i}} \operatorname{tr}\left\{ \left(\operatorname{diag}(\mathbf{p}_{2,i}) - \mathbf{p}_{2,i}\mathbf{p}_{2,i}^T \right) \left(\operatorname{diag}(\mathbf{p}_{1,i}) - \mathbf{p}_{1,i}\mathbf{p}_{1,i}^T \right) \right\}.
\end{aligned} \tag{33}$$

It follows by (20) and (32) that

$$\left| \frac{1}{n} \sum_{i=1}^n \hat{S}_i^2 - \mathbb{E}[S_i^2] \right| \to_p 0.$$

Hence, $\frac{1}{n}\sum_{i=1}^{n}(\hat{S}_i - \bar{\hat{S}})^2 \to_p \text{Var}(S_i)$.

For the numerator term, note that

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(\hat{S}_i - \bar{\hat{S}}) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(S_i - \mathbb{E}[S_i])\right| \to_p 0,$$

because: firstly,

$$\Pr\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(\hat{S}_i - S_i)\right| > t\right) \leq \frac{1}{t^2}\mathbb{E}\left[\mathbb{E}\left[\varepsilon_i^2\middle| F_i\right]\mathbb{E}\left[\hat{S}_i^2 - S_i^2\middle| F_i\right]\right] \to 0$$

by (20) and (33), mutual independence of $C_{1,i}$, $C_{2,i}$, and $(\mathbf{p}_{1,i}, \mathbf{p}_{2,i}, Y_i)$, and using finite second moment of ε_i; and, second,

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(\bar{\hat{S}} - \mathbb{E}[S_i])\right| \leq \left|\bar{\hat{S}} - \mathbb{E}[S_i]\right|\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i\right| \to_p 0$$

by the CLT and consistency of $\bar{\hat{S}}$. So by a second application of the CLT we deduce that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i(\hat{S}_i - \bar{\hat{S}}) \to_d N\left(0, \mathbb{E}\left[\varepsilon_i^2(S_i - \mathbb{E}[S_i])^2\right]\right).$$

Finally to characterize the bias term, first note that

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{S}_i - S_i)(\hat{S}_i - \bar{\hat{S}}) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{S}_i - S_i)\hat{S}_i\right| = |\bar{\hat{S}}|\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{S}_i - S_i)\right| \to_p 0$$

by consistency of $\bar{\hat{S}}$ and because

$$\mathbb{E}[(\hat{S}_i - S_i)^2] \to 0$$

holds by (20) and (33) and mutual independence of $C_{1,i}$, $C_{2,i}$, and $(\mathbf{p}_{1,i}, \mathbf{p}_{2,i}, Y_i)$. Hence by Chebyshev's inequality, we have

$$\begin{aligned}\Pr\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left((\hat{S}_i - S_i)\hat{S}_i - \mathbb{E}\left[(\hat{S}_i - S_i)\hat{S}_i\right]\right)\right| > t\right) &\leq \frac{1}{t^2}\mathbb{E}[(\hat{S}_i - S_i)^2\hat{S}_i^2] \\ &\leq \frac{1}{t^2}\mathbb{E}[(\hat{S}_i - S_i)^2] \to 0,\end{aligned}$$

where the second inequality is because $|\hat{S}_i| \leq 1$. We have therefore shown that

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{S}_i - S_i)(\hat{S}_i - \bar{\hat{S}}) - \mathbb{E}[\sqrt{n}(\hat{S}_i - S_i)\hat{S}_i] \right| \to_p 0.$$

Finally,

$$\mathbb{E}[\sqrt{n}(\hat{S}_i - S_i)\hat{S}_i] \to \kappa_1 \mathbb{E}\left[\mathbf{p}_{2,i}^T \left(\text{diag}(\mathbf{p}_{1,i}) - \mathbf{p}_{1,i}\mathbf{p}_{1,i}^T\right) \mathbf{p}_{2,i}\right] + \kappa_2 \mathbb{E}\left[\mathbf{p}_{1,i}^T \left(\text{diag}(\mathbf{p}_{2,i}) - \mathbf{p}_{2,i}\mathbf{p}_{2,i}^T\right) \mathbf{p}_{1,i}\right],$$

by (20), (33), and mutual independence of $C_{1,i}$, $C_{2,i}$, and $(\mathbf{p}_{1,i}, \mathbf{p}_{2,i}, Y_i)$. ■