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

Laura Battaglia† Timothy Christensen‡
Stephen Hansen§ Szymon Sacher¶

April 29, 2025

Abstract

Researchers now routinely use AI or other machine learning methods to estimate latent variables of economic interest, then plug-in the estimates as covariates in a regression. We show both theoretically and empirically that naively treating AI/ML-generated variables as "data" leads to biased estimates and invalid inference. To restore valid inference, we propose two methods: (1) an explicit bias correction with bias-corrected confidence intervals, and (2) joint estimation of the regression parameters and latent variables. We illustrate these ideas through applications involving label imputation, dimensionality reduction, and index construction via classification and aggregation.

JEL Codes: C11, C51, C55

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

*Authors are in alphabetical order. This paper supersedes our February 2024 working paper `https://arxiv.org/pdf/2402.15585v1`, which was circulated under a different title. SH acknowledges funding from ERC Consolidator Grant 864863, which supported his and LB's time. We thank Nick Bloom, Germain Gauthier, Evan Munro, Ashesh Rambachan, David Rossell, and Leif Thorsrud for feedback, and seminar participants at Aarhus, Bocconi, BSE, Bates, Brown, Columbia, ETH Zurich, LSE, Kent, Reserve Bank of Australia, UCSD, UPenn, USC, Warwick, the 3rd Monash-Warwick-Zurich Text-as-Data Workshop, 2024 BSE Summer Institute, 2024 FinEML Conference, 2024 UChicago Machine Learning in Economics Summer Conference, 2024 ISNPS Conference, 2024 ECONDAT Fall Meeting, and the 2024 NASM, ESIF-AIML, and ESAM Conferences of the Econometric Society. Konrad Kurczynski provided excellent research assistance.

†Department of Statistics, University of Oxford. `battaglia@stats.ox.ac.uk`

‡Department of Economics, Yale University. `timothy.christensen@yale.edu`

§Department of Economics, University College London, IFS, and CEPR. `stephen.hansen@ucl.ac.uk`

¶Graduate School of Business, Stanford University. `sacher@stanford.edu`

1 Introduction

Economists now routinely use artificial intelligence (AI) or machine learning (ML) algorithms to generate new variables. These technologies are used to quantify unstructured data such as text and images, to measure subtle concepts like uncertainty and sentiment, and to create new data sets of variables that were previously too costly, labor-intensive, or otherwise infeasible to collect.

The variables generated by AI and ML algorithms are rarely of interest in themselves, but rather are used in econometric models to address questions of cause and effect, produce forecasts, or estimate counterfactuals. In pioneering work, Baker et al. (2016) quantifies economic policy uncertainty from news text and uses it as a covariate in regressions and vector autoregressions (VARs). In more recent examples, Magnolfi et al. (2025) measures product differentiation from survey data and Compiani et al. (2025) measures product substitutability with text and images from online platforms. Both papers then use the derived measures in demand models. Gorodnichenko et al. (2023) measures tone-of-voice from audio recordings of central bank press conferences, then runs predictive regressions of financial variables on tone. Gabaix et al. (2023) imputes firm characteristics from investor holdings data and uses them to explain asset returns. Einav et al. (2024) measures patient health status from surveys then uses it in an econometric model of nursing home value added. Vafa et al. (2023) measures labor market experience from CVs and uses it to study the gender wage gap.

In standard practice, AI- or ML-generated variables are treated as regular numerical data when estimating and performing inference in downstream econometric models. We refer to this as the *two-step strategy*: variables are generated in the first step, then used as covariates in the second step. While a pragmatic initial approach, the two-step strategy has largely unknown statistical properties. One natural concern is that estimators are biased due to measurement error in the AI/ML-generated variables. Another is that inference suffers from a generated regressor problem (Pagan 1984). Conversely, results in the time-series literature suggest that plugging in estimated 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 developing reliable inference methods for working with variables generated by AI or ML, an area that is still very much in its infancy.

This paper makes two main contributions. First, we show formally that the two-step strategy can lead to invalid inference on downstream regression parameters, even in modern settings where high-performance algorithms are deployed on large data sets. Second, more

constructively, we propose two methods for valid inference: (i) bias-corrected confidence intervals, and (ii) joint estimation of the regression coefficients and latent variables. We document the performance of the methods in several empirical settings.

We consider a downstream regression where the outcome variable Y_i depends on vectors of latent variables $\boldsymbol{\theta}_i$ and observed variables $\mathbf{q}_i$. For each observation, the researcher also has an unstructured or high-dimensional data set $\mathbf{x}_i$ for estimating $\boldsymbol{\theta}_i$. To accommodate many scenarios, we stay agnostic on the form that $\mathbf{x}_i$ takes. For instance, it may be a sequence of words with textual data, an array of RGB values with image data, or a sequence of amplitudes with audio data. One scenario is *label imputation*, where $\boldsymbol{\theta}_i$ is a vector of binary labels (e.g., race or gender indicators). In this case, $\mathbf{x}_i$ (e.g., images) is used as an input to a classifier which produces predicted values $\hat{\boldsymbol{\theta}}_i$. Another scenario is *dimensionality reduction*, in which a low-dimensional representation (or embedding) $\hat{\boldsymbol{\theta}}_i$ of $\mathbf{x}_i$ is generated with an unsupervised learning model, as in Hansen et al. (2018), Bybee et al. (2024), and Ash et al. (2025) to name a few. A third scenario is *index construction*. For example, $\mathbf{x}_i$ could be a set of texts (e.g., articles, paragraphs, or sentences) which are individually classified as containing positive or negative sentiment, then aggregated and normalized to produce a sentiment score $\hat{\boldsymbol{\theta}}_i$, as in Baker et al. (2016), Caldara and Iacoviello (2022), and Gorodnichenko et al. (2023).

In the two-step strategy, the researcher first computes an estimate $\hat{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ from $\mathbf{x}_i$ for each observation, then regresses 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 context, one may wish to do inference on the coefficients of the latent or observed variables. In either case, the key question is whether this approach leads to valid inference.

To this end, we introduce an asymptotic framework in which the magnitude of the measurement error and sampling uncertainty remain comparable as the sample size increases. This framework delivers tractable approximations to the finite-sample distribution faced in practice, in which both sources of error play a role.[1] It also captures the prevailing trend of analyzing increasingly large data sets with increasingly accurate algorithms.

In this framework, we derive two new results about the two-step strategy. First, the asymptotic distribution of OLS estimators has a first-order bias due to measurement error. The bias is increasing in the scale of 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 are consistent. As a result, two-step confidence intervals have the correct width but incorrect centering,

[1] Our use of sequences of DGPs to better approximate the finite-sample behavior of estimators has a precedent in a number of contexts in economics. See, e.g., Phillips (1987), Chesher (1991), Staiger and Stock (1997), and Hahn and Kuersteiner (2002).

making them invalid for inference. This differs from a generated regressor problem, where the variance is inflated but there is no location shift. To the extent that the empirical economics literature acknowledges the two-step strategy might be a problem, concerns typically focus on standard errors. Our analysis shows these concerns are misplaced: the primary issue is bias, not incorrect standard errors.

For the case of imputed labels, the potential for AI/ML-generated variables to bias downstream estimators has been flagged in recent work, mainly in data science and political science. See Fong and Tyler (2021), Allon et al. (2023), Angelopoulos et al. (2023a,b), Zhang et al. (2023), Zrnic and Candès (2024) Miao and Lu (2024), Kluger et al. (2025) and Sanford et al. (2025) for general ML-generated variables, and Egami et al. (2023, 2024) and Ludwig et al. (2025) for variables generated by large language models.[2] These works demonstrate the inconsistency of OLS estimators in settings where the magnitude of measurement error remains fixed as the sample size increases. However, this asymptotic framework isn't necessarily appropriate in modern use cases, where high-quality algorithms are deployed on large data sets. Our analysis provides a new set of results for such cases.

Furthermore, these works propose bias corrections that require a validation sample in which both the true $\boldsymbol{\theta}_i$ and its AI/ML-generated estimate $\hat{\boldsymbol{\theta}}_i$ are observed alongside $(Y_i, \mathbf{q}_i)$.[3] The idea is to use the validation data to estimate bias, then bias-correct estimates from the main sample in which only $\hat{\boldsymbol{\theta}}_i$ is available. Such an approach is possible when the researcher can, albeit at some cost, scrutinize $\mathbf{x}_i$ and assign a ground-truth $\boldsymbol{\theta}_i$. But in that case one could simply estimate the model on the validation data alone: the AI/ML-generated data is only useful insofar as it may help improve efficiency. More problematically, $\boldsymbol{\theta}_i$ is latent in most economic use cases—for instance, one never observes true policy uncertainty, risk, or sentiment—so validation data is unavailable and these existing methods are inapplicable.

Our first inference approach is based on bias correction, but unlike existing approaches it does not require validation data. Instead, we rely on our theoretical results, which characterize the first-order asymptotic bias of OLS estimators and establish consistency of two-step standard errors. This allows us to perform an analytical bias correction, then re-center the usual confidence intervals at the bias-corrected estimator to perform valid inference. Our bias corrections are general and widely applicable. We specialize them to AI/ML-generated binary labels and dimension reduction. For the former, bias-correction can be performed without validation data provided one has a measure of the classifier's expected false-positive rate.[4] For the latter, bias correction can be performed using the estimated low-dimensional

[2]There is also recent work in economics that considers the complementary problem of imputed dependent variables as opposed to imputed covariates; see Rambachan et al. (2025) and Modarressi et al. (2025).

[3]This approach is related to an older literature on estimation with auxiliary data (Chen et al. 2008).

[4]The expected false-positive rate may be estimated from a validation sample, but it may also be available

representation.

It is important to note that the measurement error in AI/ML generated variables may be "nonclassical" (i.e., correlated with the true latent $\boldsymbol{\theta}_i$). This makes it difficult for researchers to know even the sign of the bias ex ante: there may be attenuation or amplification. Indeed, Section 6.3 shows that bias can be positive or negative in the case of index construction. Nonclassical measurement error is also much more difficult to correct for than classical measurement error, requiring specialized methods—see, e.g., Schennach (2022) for a discussion.

Our bias corrections are convenient to apply, but they may not be available for all types of AI or ML algorithms. They also rely on the magnitudes of measurement error and sampling uncertainty being comparable. To perform inference without validation data in settings where this is not the case, we introduce a second approach based on joint maximum likelihood estimation of the models for latent variable estimation and regression. In this approach, the model linking $\mathbf{x}_i$ with the latent $\boldsymbol{\theta}_i$ is analogous to an "observation equation" in state-space models. This requires some more careful modeling of how $\boldsymbol{\theta}_i$ and $\mathbf{x}_i$ are related, but we demonstrate its feasibility with three distinct and non-exhaustive applications: AI/ML generated binary labels, dimension reduction, and AI/ML generated indices.

While joint estimation is straightforward in theory, it presents a computational challenge due to the large number of latent $\boldsymbol{\theta}_i$ that must be integrated out of the likelihood. To address this, we use Hamiltonian Monte Carlo, 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.[5]

We introduce three applications to illustrate the theoretical results. The first illustrates label imputation. Hansen et al. (2023) uses a Large Language Model to classify each job posting in the Lightcast dataset as offering remote work or not. These imputed labels can be merged with other posting-level metadata to study the causes and consequence of remote work adoption. We focus on the relationship between wage inequality and remote work by regressing the posted wage on the remote indicator and controls. The classifier achieves a high test-set accuracy of 99%, so one might expect that measurement error is inconsequential. But our theory shows that the important quantity is measurement error *relative* to sampling uncertainty, and the Lightcast dataset has hundreds of millions of individual observations.

externally. To give a recent example, Bursztyn et al. (2024) uses a ML algorithm to classify charitable donors' names by ethnicity. They estimate the accuracy of the classifier using an external sample of North Carolina voter registration data which contains self-reported ethnicity (but not data on donations or other controls).

[5]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 custom code to estimate, which makes adapting the model difficult for non-specialists.

We show via a case study that bias correction and joint estimation both estimate notably stronger effects of remote work on posted wages than the two-step strategy.

The second application illustrates regressors derived from dimension-reduction algorithms. Bandiera et al. (2020) conducts a time-use survey to document behavioral differences among CEOs and their impact on firm performance. The authors use latent Dirichlet allocation (Blei et al. 2003)—a factor model for discrete data—to represent CEO time-use behavior in a low-dimensional space. This representation is then included as a covariate along with other firm controls in a sales regression. We replicate this two-step strategy and find the estimated impact of behavior on performance aligns with estimates obtained via bias correction and the joint estimation strategy. Our theory predicts this will hold when measurement error is low compared to sampling uncertainty. We then re-estimate the model using a 10% subsample of time units for each CEO to scale-up measurement error. Here we find the two-step strategy produces insignificant behavioral effects, while both corrections produce significant effects.

The third application illustrates index construction via classification and aggregation. Central bank communication has become a major research and policy topic over the past decade, and linking market reactions to communication often involves quantifying the latter from unstructured data. We replicate the hawkish sentiment measure from Gorodnichenko et al. (2023) which classifies individual paragraphs of FOMC statements as hawkish or dovish. Paragraphs are then aggregated to form a meeting-level share which proxies continuous, latent sentiment. Following the two-step strategy, we regress the path factor (Gürkaynak et al. 2005)—a measure of movement in the long end of the yield curve—on sentiment and find weakly positive effects. With joint estimation, however, the estimated effect size and R^2 are both nearly three times larger, which shows the value of our correction for prediction as well as for inference.

Finally, in simulation exercises calibrated to the empirical applications, we find that the two-step strategy performs poorly—in terms of bias in estimated coefficients and coverage of confidence intervals—relative to both bias correction and joint estimation. This provides further evidence that measurement error distorts inference in the two-step strategy, while our corrections reduce bias and restore valid inference even in challenging empirical settings.

Our overall message is that the increasingly common practice of using regressors generated by AI or ML can lead to invalid inference, but practical solutions exist. We view our proposed bias correction and joint estimation approaches as robust, widely applicable starting points for empirical analysis. For instance, an emerging line of research uses text-derived sentiment indices as inputs into forecasting models. Our analysis can be extended to show how errors in these indices lead to biased forecasts, while our solutions can improve the performance of these forecasting methods. Likewise, the industrial organization literature increasingly uses

embedded representations of firms and products to model market behavior and demand. Our solutions can be adapted to these settings as well. Going forward, it is important to establish which algorithms and econometric models are most susceptible to measurement error and associated inference problems. More generally, inference problems arising from the use of AI/ML-generated variables should more widely recognized in order to fully harness the potential of AI/ML methods in empirical economics.

The rest of the paper proceeds as follows. Section 2 provides a simple setting illustrating why the two-step strategy leads to biased inference and how our proposed solutions can help. Section 3 introduces the more general framework and presents three empirical applications. Sections 4 and 5 present, respectively, the main theoretical analysis of the two-step strategy and the proposed solutions. Section 6 presents simulation results and Section 7 concludes.

2 A Simple Example

This section presents a simple model to illustrate how the two-step strategy leads to biased inference, and how our proposed methods can restore valid inference.

2.1 Model

The model is loosely based on Baker et al. (2016). Suppose we are interested in the effect γ of θ_i (policy uncertainty in month i) on Y_i (employment or investment, say, in month $i+1$) in the regression model

$$Y_i = \alpha + \gamma\theta_i + \varepsilon_i. \tag{1}$$

Policy uncertainty is a nebulous concept that is difficult to precisely define let alone observe. Baker et al. (2016) forms EPU indices from monthly counts of articles in ten newspapers containing certain terms, which they convert to an index. Evidently there is measurement error due to the sampling of articles: one could change the set of newspapers surveyed and obtain a quantitatively different (but related) measure.[6] To capture this, consider

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

where C_i is the number of articles sampled in month i, X_i is the number of these that relate to uncertainty, and θ_i is true policy uncertainty. We observe X_i, Y_i, and C_i but not θ_i. One can estimate θ_i using $\hat{\theta}_i = X_i/C_i$, as done by Baker et al. (2016, p. 1599).

[6] Misclassification of articles is a second source of measurement error. We sidestep this for now for sake of exposition, but account for it in later sections.

2.2 Problem with the Two-Step Strategy

In this example, the two-step strategy computes the OLS estimate $\hat{\gamma}$ from regressing Y_i on $\hat{\theta}_i$, then performs standard OLS inference for γ. This approach ignores the fact that $\hat{\theta}_i$ is a noisy estimate of θ_i, potentially leading to biased estimates and invalid inference.

We use asymptotics to tractably approximate the finite-sample problem faced by the researcher, where $\hat{\gamma}$ is computed from $(Y_i, X_i, C_i)_{i=1}^n$. Both measurement error and sampling error affect the properties of $\hat{\gamma}$ in finite samples. We therefore consider a sequence of populations indexed by the sample size n, where 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

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

In this sequence of DGPs, the variance of $\hat{\theta}_i$, which is proportional to C_i^{-1}, is of the same order of magnitude as sampling uncertainty. The parameter κ controls the relative importance of measurement error, with larger values of κ giving relatively greater importance to measurement error. Working with this sequence of DGPs therefore allows us to gain insights about how $\hat{\gamma}$ behaves when both measurement and sampling error are present.

Under suitable regularity conditions (see Theorem 1), one can show that

$$\begin{aligned} &\sqrt{n}(\hat{\gamma} - \gamma) \to_d N\left(-\kappa\,\gamma \frac{\mathbb{E}[\theta_i(1-\theta_i)]}{\mathrm{Var}(\theta_i)}, \frac{\mathbb{E}[\varepsilon_i^2(\theta_i - \mathbb{E}[\theta_i])^2]}{\mathrm{Var}(\theta_i)^2}\right), \\ &\frac{\sum_{i=1}^n \hat{\epsilon}_i^2(\hat{\theta}_i - \bar{\theta}_n)^2}{\sum_{i=1}^n (\hat{\theta}_i - \bar{\theta}_n)^2} \to_p \frac{\mathbb{E}[\varepsilon_i^2(\theta_i - \mathbb{E}[\theta_i])^2]}{\mathrm{Var}(\theta_i)^2}, \end{aligned} \tag{4}$$

where $\bar{\theta}_n$ is the sample mean of $\hat{\theta}_i$ and $\hat{\epsilon}_i$ is the OLS residual. The first result shows $\hat{\gamma}$ is asymptotically normally distributed with the same variance as if Y_i was regressed on the true latent θ_i, but with a centering that differs from zero when $\kappa > 0$. The second result shows OLS standard errors are consistent, irrespective of κ. Taken together, these results imply that two-step confidence intervals (given by $\hat{\gamma} \pm 1.96$ times the OLS standard error) have the correct width, but incorrect centering whenever $\kappa > 0$. Moreover, the asymptotic bias of $\hat{\gamma}$, and therefore the degree of under-coverage of two-step CIs, is increasing in κ.

2.3 Proposed Solutions

2.3.1 Bias Correction

Our first proposed solution is a straightforward bias correction. This approach simply constructs an estimate of the bias and adds it back to the two-step estimator $\hat{\gamma}$. The bias correction follows easily from (4), so the bias corrected estimator $\hat{\gamma}^{bc}$ is

$$\hat{\gamma}^{bc} = \left(1 + \frac{\hat{\kappa}}{\sqrt{n}} \frac{\sum_{i=1}^{n} \hat{\theta}_i(1-\hat{\theta}_i)}{\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta}_n)^2}\right)\hat{\gamma},$$

where $\hat{\kappa} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} C_i^{-1}$. Bias corrected confidence intervals are then simply $\hat{\gamma}^{bc} \pm 1.96$ times the OLS standard error. See Theorem 4 for a formal justification for this approach.

2.3.2 Joint Estimation

A second approach is joint maximum likelihood estimation of (1) and (2). This approach treats (2) analogously to an observation equation in a state-space model, with θ_i as a latent variable.

We start by assuming the error terms in (1) have probability density function $\sigma^{-1}f(\varepsilon/\sigma)$. Combining with (2), this yields the likelihood

$$f(Y_i, X_i | C_i, \theta_i; (\gamma, \alpha, \sigma)) \propto \frac{1}{\sigma} f\left(\frac{Y_i - \alpha - \gamma\theta_i}{\sigma}\right) (\theta_i)^{X_i}(1-\theta_i)^{C_i - X_i}.$$

Since θ_i is latent we proceed in the spirit of random effects and assume θ_i is drawn from a distribution with probability density function g on $[0,1]$.[7] Note that the effect of this prior will be dominated by the data as C_i becomes large. We then integrate out θ_i to produce a likelihood in terms of the observed data:

$$f(Y_i, X_i | C_i; (\gamma, \alpha, \sigma)) = \int_0^1 f(Y_i, X_i | C_i, \theta_i; (\gamma, \alpha, \sigma)) g(\theta_i)\, d\theta_i.$$

We estimate γ by maximizing the log-likelihood

$$L_n((\gamma, \alpha, \sigma)) = \sum_{i=1}^{n} \log f(Y_i, X_i | C_i; (\gamma, \alpha, \sigma)).$$

Inference is performed using standard asymptotics for maximum likelihood estimators.

[7] One can easily allow the distribution of θ_i to depend on covariates, as in correlated random effects. We suppress this for now, but adopt such an approach in the empirical applications.

3 General Setup and Applications

In the general model, we wish to estimate and perform inference on the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ of the linear regression model

$$Y_i = \boldsymbol{\gamma}^T \boldsymbol{\theta}_i + \boldsymbol{\alpha}^T \mathbf{q}_i + \varepsilon_i, \tag{5}$$

where $\boldsymbol{\theta}_i$ is now extended to be a vector of *latent* variables of interest, $\mathbf{q}_i$ is a vector of *observed* quantitative variables, and $\mathbb{E}\left[\varepsilon_i(\boldsymbol{\theta}_i, \mathbf{q}_i)\right] = 0$. For each observation i we also have unstructured or high-dimensional data $\mathbf{x}_i$, from which an estimate $\hat{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ can be derived. Thus, the researcher's dataset is a random sample $(Y_i, \mathbf{q}_i, \mathbf{x}_i)_{i=1}^n$. The parameter $\boldsymbol{\gamma}$ is typically the key object of interest, but in some cases (e.g., Avivi 2024) $\boldsymbol{\alpha}$ may be the focus, with $\boldsymbol{\theta}_i$ serving as a control variable derived from unstructured data.

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

(i) Compute estimates $\hat{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ for all observations $i = 1, \ldots, n$.

(ii) Regress Y_i on $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{q}_i$. Compute standard errors and confidence intervals, 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 proxies for the true latent covariates $\boldsymbol{\theta}_i$ in (5). Step (ii) overlooks this issue and treats the estimates $\hat{\boldsymbol{\theta}}_i$ as regular numeric data. This raises the possibility that two-step estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are biased. Moreover, conventional standard errors and confidence intervals do not account for the additional variation arising 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, in Section 4 we shall analyze the two-step strategy and formally demonstrate why it can lead to biased estimates and inference. Many specific cases can be captured by this setup. We consider three here.[8]

Application 1: AI/ML-Generated Labels. Economists now routinely use AI or ML methods to impute missing covariates. A leading use case involves regressions of an outcome Y_i on a latent binary variable θ_i (e.g., indicating positive/negative sentiment of a news article or racial group membership) and observed controls $\mathbf{q}_i$.[9] Examples include Goldsmith-Pinkham and Shue (2023), Adams-Prassl et al. (2023), Argyle et al. (2025), and Wu and Yang (2024). Unstructured data $\mathbf{x}_i$ (e.g., article text or voter registration data) is often used to predict θ_i using a classification algorithm. The two-step strategy entails first generating

[8] The previous version of the paper considers further extensions, for example regression onto similarity measures between vector representations of unstructured data.

[9] We present the case of scalar $\boldsymbol{\theta}_i$ in the main text and defer the case of multiple categories to Appendix B.

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. Although sophisticated modern classifiers have low error rates, they are often used to impute missing observations for large data sets. As a result, measurement error from misclassification may be non-negligible relative to sampling error in the downstream regression, invalidating two-step inference.

Application 2: Topic Models. A large empirical literature uses topic models to reduce the dimension of unstructured data. Examples include Hansen et al. (2018), Mueller and Rauh (2018), Larsen and Thorsrud (2019), Thorsrud (2020), Adams et al. (2021), Bybee et al. (2024), and Ash et al. (2025) with text, Draca and Schwarz (2021), and Munro and Ng (2022) with survey data, and Nimczik (2017) and Olivella et al. (2021) with network data.

Here $\mathbf{x}_i = (x_{i,j})_{j=1}^V$ is a V-dimensional vector of feature counts, with $x_{i,j}$ counting the number of times feature j appears in observation i. For instance, in text applications, $x_{i,j}$ counts the number of times word j appears in document i. The vector $\mathbf{x}_i$ follows a Multinomial distribution with a factor structure. There are $K < V$ distributions $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K \in \Delta^{V-1}$, the $(V-1)$-dimensional simplex. Each $\boldsymbol{\beta}_k$ represents a common factor (or "topic"). Each observation i is characterized by a latent vector $\boldsymbol{w}_i \in \Delta^{K-1}$. The elements $w_{i,k}$ of $\boldsymbol{w}_i$ represent the weight of $\boldsymbol{\beta}_k$ in generating $\mathbf{x}_i$. The count probabilities for observation i are $\mathbf{p}_i = \sum_{k=1}^K \boldsymbol{\beta}_k w_{i,k} = \mathbf{B}^T \boldsymbol{w}_i$, where $\mathbf{B}^T = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]$. Combining these elements yields

$$\mathbf{x}_i | (C_i, \boldsymbol{w}_i) \sim \text{Multinomial}(C_i, \mathbf{B}^T \boldsymbol{w}_i), \tag{6}$$

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

In the two-step approach, $\mathbf{B}$ and $(\boldsymbol{\theta}_i)_{i=1}^n$ are estimated using Latent Dirichlet Allocation (Blei et al. 2003, LDA) or more recent methods (e.g., Bing et al. (2020), Wu et al. (2023), Ke and Wang (2022)), then Y_i is regressed on $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{q}_i$. The source of measurement error is sampling error in the estimated topic weights $(\hat{\boldsymbol{\theta}}_i)_{i=1}^n$, which is proportional to C_i^{-1}. The quantity $\mathbb{E}[C_i^{-1}]$ controls the overall rate of measurement error.

Application 3: AI/ML-Generated Indices. A third use case involves constructing indices by classification and aggregation. For instance, Baker et al. (2016) constructs policy uncertainty indices by first classifying news articles based on whether they relate to policy uncertainty, and then aggregating the results over time into monthly or quarterly indices. See also Caldara and Iacoviello (2022) and Gorodnichenko et al. (2023), among others.

Extending the simple example of Section 2, let C_i denote the number of articles to be classified in month i, and let N_i represent the number of articles classified as pertaining

to policy uncertainty. The quantity $\hat{\theta}_i = N_i/C_i$ is a natural measure of the true latent uncertainty $\theta_i \in [0,1]$, where $\theta_i = 0$ indicates no uncertainty and $\theta_i = 1$ indicates maximal uncertainty. As each individual article classification may be subject to some error, there are now two sources of measurement error: misclassification error and sampling uncertainty (the set of articles are a sample from the broader corpus of news). Both can be accounted for using a topic model. Suppose the misclassification rates are constant across observations. Then $\mathbf{n}_i = (N_i, C_i - N_i)^T$ follows the distribution in (6), 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 relating to uncertainty is correctly classified, and β_0 is the probability that an article not relating to uncertainty is misclassified.

As in the simple example in Section 2, we consider two strategies to correct bias and restore valid inference. The first strategy involves bias-corrected estimators and confidence sets, which we formally develop and provide theoretical justification for in Section 5.1. The second is joint estimation, the implementation details of which are discussed in Section 5.2. In the remainder of the section, we illustrate the problems with two-step estimation—and how the proposed solutions fix them—across three applications drawn from diverse empirical literatures. Each application corresponds to one of the examples above.

3.1 Remote Work and Wage Inequality

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 metadata on occupation, firm, location, job title, and the posted wage, and a textual description of the job. Hansen et al. (2023) uses the posting text to impute a binary label indicating whether or not the posting offers remote work. The authors collect human labels from Amazon Mechanical Turk for a sample of postings and use them to fine-tune DistilBERT (Sanh et al. 2020), a Large Language Model. The classifier achieves high overall accuracy in a held-out test set (98%-99%). The authors then impute a remote work label for every posting in the corpus.

The data can be used to answer numerous questions about the causes and consequences of remote work. One important question is the degree of wage inequality in remote work

Table 1: Estimates of Impact of Remote Work on Posted Wages (San Diego, NAICS 72)

Controls	Estimation Strategy		
	Two-Step	Bias Correction	Joint
None	0.648 [0.600, 0.697]	1.052 [0.778, 1.327]	0.563 [0.532, 0.595]
SOC2 effects	0.364 [0.322, 0.406]	0.641 [0.446, 0.836]	0.448 [0.415, 0.480]

Note: Point estimates and 95% confidence intervals for the slope coefficient in a regression of posted log salary on a binary remote work indicator, with and without occupation fixed effects at the SOC2 level. The sample consists of 16,315 job postings for 2022 and 2023 with "San Diego, CA" recorded as the city and "72" recorded as the NAICS2 industry code of the advertising firm.

arrangements, which the data has previously been used to study (Lambert et al. 2023). To explore this issue, we use regression (5) taking Y_i as the log of the advertised wage for posting i.[10] Here $\theta_i \in \{0, 1\}$ is a latent indicator of whether posting i offers remote work. The two-step approach replaces the latent θ_i with the predicted label $\hat{\theta}_i \in \{0, 1\}$ from the classifier. Given the remote work classifier achieves high test-set accuracy, one may believe that such classification error is unlikely to affect inference meaningfully in this application. However, what matters is not measurement error *per se* but its magnitude relative to sampling error. In regressions with a large number of observations, sampling error is potentially small enough that even small classification errors can distort inference.

To illustrate the impact of classification error, we begin by organizing the data into NAICS2 industry $\times$ city cells for all job postings in 2022 and 2023 in the United States. These years coincide with peak post-pandemic incidence of remote work. Our theory in Section 4 shows in which situations classification error is likely to produce the largest distortions, and we use it to select for case study NAICS 72 industry (Accommodation and Food Services) postings in San Diego. After removing internships and observations with missing salary data, there are 16,315 observations. The first column of Table 1 contains two-step estimates of γ. In the initial regression model, we include no controls and find an estimated 65 log point effect on wages of remote work. Since this may in part reflect occupation composition, we next include SOC2 fixed effects in $\mathbf{q}_i$, which reduces the estimated γ to 36 log points.

The positive association between remote work and posted wages is consistent with the descriptive evidence in Lambert et al. (2023). Nevertheless, its strength may be distorted in the two-step approach. A key quantity governing the bias is the expected false positive rate (FPR), which can be estimated by reading a random sample of postings and counting

[10] We form the advertised wage by averaging the *salary_from* and *salary_to* fields in the Lightcast data.

the number that were mis-classified as positive. In this example, we took a random sample of 1000 postings, read the 26 that were classified as positive, and found nine were classified incorrectly, so the expected FPR is 0.009. Despite the small FPR, bias correction produces much larger effects: with no controls, the effect increases by 62% (to 105 log points) and, with controls, by 76% (to 64 log points).[11] In many economic applications of AI and ML methods, classification accuracy is well below its level here. Finding large effects even in this setting shows that classification error can be of first-order importance.

In addition, we formulate a joint model over θ_i, $\hat{\theta}_i$, and Y_i. While this produces a smaller estimated γ without controls, in the preferred specification with occupation effects we continue to find a larger effect that falls outside the two-step confidence intervals. One reason why joint estimation and bias correction may not coincide is that human labels themselves may not represent the ground truth even in the best-designed audit. For example, the 2025 Economic Report of the President (Council of Economic Advisers 2025) discusses strengths and weaknesses of the Hansen et al. (2023) database and points out that not every posting will explicitly mention remote work even when the firm in practice offers it.[12]

3.2 CEO Time Use and Firm Performance

The role of CEOs in shaping firm performance is important for many academic and policy debates, but until recently little data existed on what CEOs do with their time. To fill this evidence gap, Bandiera et al. (2020) collects and analyzes survey data on CEO time use in a sample of manufacturing firms. The paper describes salient differences in executive time use and relates those differences to firm and CEO characteristics and firm outcomes.

The survey consists of five questions with categorical responses: (Q1) the type of activity (meeting, public event, etc.); (Q2) duration of activity (15m, 30m, etc.); (Q3) whether the activity is planned or unplanned; (Q4) the number of participants in the activity; (Q5) the functions of the participants in the activity (HR, finance, suppliers, etc.). Survey responses are recorded for each 15-minute interval of a given week, e.g. Monday 8am-8:15am, Monday 8:15am-8:30am, and so forth. The sample consists of 916 CEOs.

The data are modeled as a topic model, with $V = 654$ answer combinations observed across the five questions, and $x_{i,j}$ denoting the number of times combination j appears in

[11] Bias-corrected confidence intervals are wider to account for the uncertainty in the estimated FPR.

[12] The relevant passage is

> Job openings data can also shed light on whether remote work is here to stay. While the information can be murky—given that not every hybrid or remote job advertises itself as such, and the tendency to mention remote work in job postings may change over time—examining recent trends is useful (Council of Economic Advisers 2025).

Table 2: Estimates of Impact of CEO Behavior on Firm Performance

Sample	Estimation Strategy		
	Two-Step	Bias Correction	Joint
Full	0.405 [0.224, 0.585]	0.474 [0.294, 0.655]	0.402 [0.240, 0.603]
10% Subsample	0.227 [-0.038, 0.492]	1.054 [0.789, 1.319]	0.439 [0.153, 0.711]

Note: The first row presents point estimates and 95% confidence intervals for a regression of log sales on the CEO behavior index from the replication data of Bandiera et al. (2020). The second row presents estimates and confidence intervals using a 10% subsample of time use responses per CEO.

the diary of CEO i. The authors reduce the dimensionality of the feature space using LDA with $K = 2$ topics. The estimated $\hat{\boldsymbol{\beta}}_1$ places relatively higher mass on features associated with "management" like visiting production sites and one-on-one meetings with employees or suppliers, while $\hat{\boldsymbol{\beta}}_2$ places relatively higher mass on features associated with "leadership" like communicating with other C-suite executives and holding large, multi-function meetings. The leadership weight $\hat{\theta}_i$—which the authors call a "behavior index"—is thus a measure of the tendency of CEO i to engage in leadership activities. Bandiera et al. (2020) regress log sales on $\hat{\theta}_i$ and firm controls, finding a positive association between leadership and firm performance. The paper's use of the two-step strategy may, however, lead to invalid inference.

To explore this possibility, we first replicate the authors' two-step strategy, estimating θ_i by LDA, then regressing the log sales of each CEO's firm on $\hat{\theta}_i$ and controls $\mathbf{q}_i$ including log employment, country fixed effects, and survey-wave fixed effects. Results reported in Table 2 show that, according to the two-step approach, moving from a CEO who only spends time in management to one who only spends time in leadership is associated with a 40 log point increase in sales. What's more, neither bias correction nor joint estimation shifts the estimated effect size by a large amount. This shows that bias is not an inevitable part of using AI/ML-generated variables, but rather depends on the empirical setting.

One reason why measurement error is less important here lies in the κ expression (3) for the simple model. While not directly applicable here as the topic model structure is more complex, it still allows one to qualitatively compare sampling error (reflected by $\sqrt{n}$) and measurement error (reflected by $\mathbb{E}\left[C_i^{-1}\right]$). The empirical analogue of this expression is 0.44. In other words, there are a relatively large number of survey responses per CEO compared to the number of surveyed CEOs, meaning the measurement error in $\hat{\theta}_i$ is small relative to sampling error. To increase measurement error, we take a random 10% subsample of survey responses for each CEO, which can be thought of as observing half a day of behavior rather

than a five-day workweek. The two-step approach now estimates a smaller insignificant effect, while both bias correction and joint estimation continue to estimate a significant effect.

3.3 Central Bank Communication

Public communication is an increasingly important tool for central banks (Blinder et al. 2008, Blinder 2018) and disentangling its effects is an ongoing research challenge. Communication events are often accompanied by rich, unstructured data like the content of speeches and press conferences, which require some quantification prior to econometric analysis. To illustrate this, we consider market reactions to FOMC policy announcements. Using high-frequency yield curve movements around FOMC announcements, Gürkaynak et al. (2005) extracts separate "target" and "path" factors which account for large shares of observed variation in short- and long-run yields, respectively. One view is that policy rate news drives short rates (i.e., the target factor) while communication in the form of written statements drives long rates (i.e., the path factor) by shifting expectations of future actions.

To test this, we estimate (5) with Y_i as the path factor for FOMC meeting i (as updated in Acosta et al. 2024), θ_i as the hawkish sentiment of the FOMC written statement, and $\mathbf{q}_i$ as a constant and the shadow short rate (Wu and Xia 2016), which accounts for policy variation during the zero-lower-bound period. The sample period is Feb 1995 through June 2023, during which 200 FOMC meetings take place.[13]

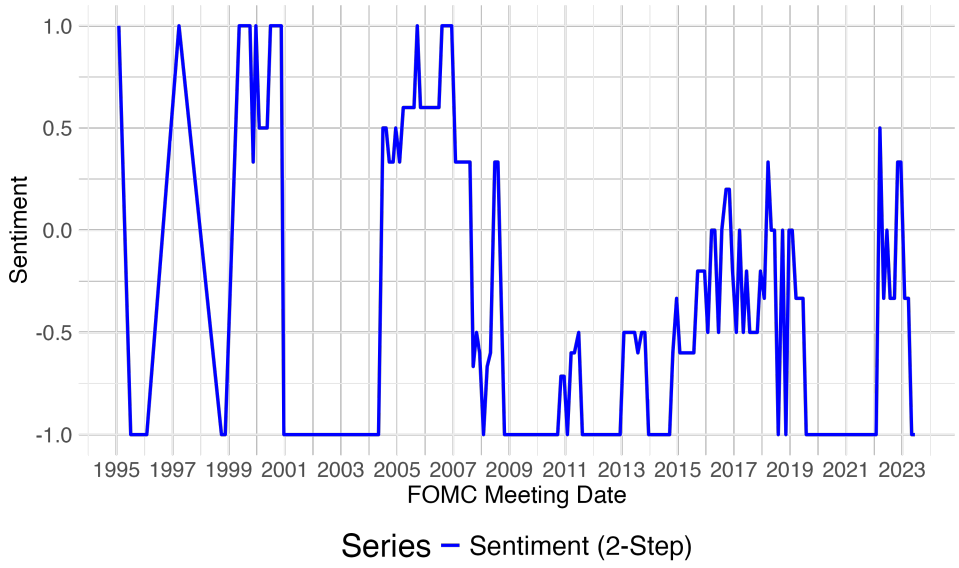
Of course, hawkish sentiment θ_i is latent. To estimate it, we replicate the approach of Gorodnichenko et al. (2023) which uses classification and aggregation to build a sentiment index.[14] Gorodnichenko et al. (2023) provides 1,243 sentences with human labels from FOMC statements from 1997 through 2010. Each sentence assigned to one of three mutually exclusive categories: *hawkish* (243 sentences), *dovish* (511), or *neutral* (489). We split the sentences into 1,118 training observations and 125 test observations and fine-tune BERT (Devlin et al. 2019)—a well-known Large Language Model—on the training data for label prediction.[15] We assign each sentence in the test data to its most likely class based on the trained model and obtain test-set accuracy of 0.85, slightly higher than in the original paper (0.81). We then predict a label for all paragraphs from FOMC statements in our sample.[16]

[13] The sample is determined by the availability of the path factor and shadow short rate data.

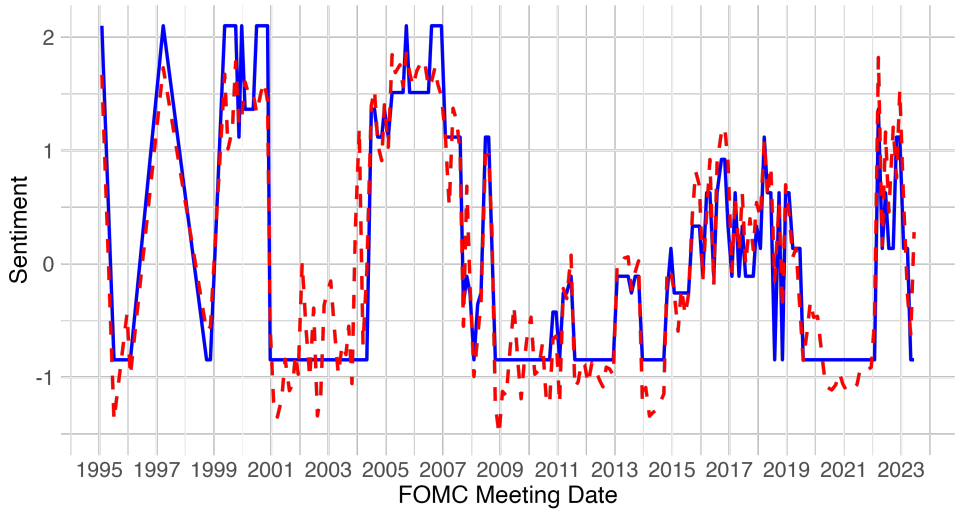
[14] Gorodnichenko et al. (2023) studies the market impact of the tone of voice of FOMC Chairs during post-meeting press conferences, and uses the hawkish sentiment of the written statement as a control.

[15] More precisely, we fine tune `bert-base-uncased` via the Hugging Face library in Python.

[16] Here we follow Gorodnichenko et al. (2023) and classify individual *paragraphs* in the full sample although BERT is trained on human-labeled *sentences*. The optimal level of document granularity for classification prior to aggregation is an open question. Gorodnichenko et al. (2023) classifies individual paragraphs of FOMC press conferences in addition to FOMC statements, but press conferences only began in 2011 where statements have been released since the mid-1990s.



(a) Unscaled Series



(b) Scaled Series

Figure 1: Time Series of FOMC Statement Sentiment

Note: The left panel plots two-step sentiment $\hat{\theta}_i$. The right panel plots standardized series for $\hat{\theta}_i$ and estimated sentiment from the joint model. The sample period covers 200 FOMC meetings from Feb 1995 through June 2023.

Table 3: Impact of FOMC Statement Sentiment on Longer-Term Yields

	Estimation Strategy	
	Two-Step	Joint
Sentiment (θ_i)	0.038	0.114
	[0.005, 0.071]	[0.027, 0.198]
Policy Rate (q_i)	-0.004	-0.003
	[-0.013, 0.004]	[-0.011, 0.004]
β_0		0.009
		[0.001, 0.026]
β_1		0.676
		[0.585, 0.768]
R^2	0.0425	0.1429

Note: The first column reports point estimates and 95% confidence intervals using the two-step approach for estimating the relationship between hawkish sentiment in FOMC statements and the path factor of Gürkaynak et al. (2005) updated by Acosta et al. (2024). The second column reports results from joint estimation.

The overall sentiment measure for FOMC meeting i is $\hat{\theta}_i = N_i/C_i$ where C_i is the number of sentences classified as hawkish or dovish, and N_i is the number classified as hawkish. Figure 1a shows the evolution of this measure over the sample period.

The first column of Table 3 contains the results of two-step estimation. There is some evidence consistent with communication rather than policy driving the path factor. The coefficient on $\hat{\theta}_i$ is positive and significant while that on the policy rate is essentially zero. At the same, the effect is somewhat weak. The standard deviation of the path factor in the sample is 0.103, so the estimated $\hat{\gamma}$ implies a one-third standard deviation effect of θ_i moving from 0 to 1. The regression also has a low R^2.

We do not develop a bias correction for index construction.[17] We do however implement joint estimation. The second column of Table 3 shows the results. The effect size on sentiment nearly triples, as does the R^2 value. Confidence intervals are a little wider using this method, as it accounts for the fact that some FOMC announcements are relatively short, making their corresponding θ_i values difficult to infer. Figure 1b shows that estimated sentiment from the two approaches co-move strongly, but estimated sentiment from the joint model is smoother.[18] In short, we find results consistent with measurement error in the two-step

[17]Our bias corrections are valid provided measurement error and sampling error are comparable. The relatively short length of FOMC announcements and large misclassification errors we report below suggests measurement error may be sizable in this case. Joint estimation remains valid in such settings.

[18]While the joint approach is based on maximizing the integrated likelihood, our inference algorithm is based on sampling from the posterior distribution over θ_i and regression parameters. Sentiment from the joint model is the average value of θ_i draws.

sentiment estimate which weakens the estimated market reaction to FOMC statements.

These results also have implications for the large literature that regresses outcomes on sentiment. Shapiro et al. (2022) compares a variety of classifiers for predicting human labels for news article sentiment, including BERT, and reports substantial error even for the highest-performing methods (see Tables 1-3 in the paper). Approaches like ours can restore valid inference in such settings.

Table 3 also shows estimates of β_0 and β_1, and shows the latter is well below that implied by the classifier. Economic theory predicts that the difference between realized and expected hawkishness is the relevant object for generating market reactions, whereas the human labels only classify realized hawkishness. While a full analysis is outside the scope of the paper, joint estimation in principle allows β_0 and β_1 to adjust to account for this broader notion of misclassification.

4 Why Two-Step Inference is Biased

The empirical applications show that measurement error in AI/ML-generated variables can bias inference in a variety of settings. Each application has a particular algorithm for estimating $\hat{\boldsymbol{\theta}}_i$ that differs in important respects: supervised vs unsupervised learning, whether $\hat{\boldsymbol{\theta}}_i$ is the aggregation of multiple or single classified documents, and so forth. Further varieties are also common in the literature. To better understand the biases arising from measurement error and how to fix them, we begin by abstracting away from algorithmic-specific details of how $\hat{\boldsymbol{\theta}}_i$ is estimated. This allows us to develop a general, broadly applicable framework, which we then tailor to some specific use cases.

4.1 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

$$\hat{\boldsymbol{\psi}} = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i Y_i \right). \tag{7}$$

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

We use asymptotic theory to derive tractable approximations to the finite-sample distribution of $\hat{\boldsymbol{\psi}}$. Both measurement error in $\hat{\boldsymbol{\theta}}_i$ and downstream sampling error determine the finite-sample distribution. To ensure that asymptotics deliver a useful approximation, we adopt a framework in which both sources of error remain present as the sample size n becomes large. We do so be allowing the precision of $\hat{\boldsymbol{\theta}}_i$ to increase with n at an appropriate rate. Formally, we consider a sequence of populations in which the distribution of $(Y_i, \boldsymbol{\theta}_i, \mathbf{q}_i)$ is held fixed and the conditional distribution of $\hat{\boldsymbol{\theta}}_i$ given $(Y_i, \boldsymbol{\theta}_i, \mathbf{q}_i)$ varies with n, so that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\boldsymbol{\theta}}_i (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T \to_p \kappa \, \boldsymbol{\Omega} \tag{8}$$

as $n \to \infty$, where $\boldsymbol{\Omega}$ is a finite non-random matrix and κ is a non-negative constant. The constant κ represents the *relative* magnitudes of measurement error and sampling error. The matrix $\boldsymbol{\Omega}$ is related to the variance of measurement error in settings where the error in $\hat{\boldsymbol{\theta}}_i$ is "classical". However, condition (8) also allows for "non-classical" measurement error, which is needed to accommodate a number of important use cases—including imputed categorical labels (Aigner 1973). We show below that (8) 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.

We view this asymptotic framework as appropriate for approximating modern use cases where high-performance algorithms are deployed on large data sets.[19] While it is not meant to be taken literally, a heuristic interpretation is that it captures the prevailing trend whereby increasingly large datasets are analyzed by increasingly accurate algorithms.

Having introduced the asymptotic framework, we now introduce the assumptions. We first present a general "high-level" set of assumptions, then verify them below within the context of the running examples. In what follows, notions of convergence in probability and distribution should be understood as holding along this sequence of populations satisfying condition (8). Let $\hat{\varepsilon}_i = Y_i - \hat{\boldsymbol{\psi}}^T \hat{\boldsymbol{\xi}}_i$. Throughout, we let $\mathbf{0}$ denote a matrix or vector of zeros whose dimension is determined by the context.

Assumption 1. *(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) $\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T \to_p \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]$, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{\boldsymbol{\theta}}_i (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T \to_p \kappa \, \boldsymbol{\Omega}$, *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* $n \to \infty$.

[19] In such scenarios, conventional measurement error frameworks in which the size of the measurement error remains fixed seem inappropriate, since measurement error bias eventually dominates and estimators are inconsistent. For completeness, we provide a set of results for this case in Appendix B.

Assumption 1(i) is standard. Assumption 1(ii) can be verified using appropriate laws of large numbers. The first part of Assumption 1(iii) imposes a standard CLT condition while the second part is only used to establish consistency of standard errors. We focus on the case where the data are independent and identically distributed to simplify exposition, though this can be relaxed and the results can easily be extended to general types of dependence.

We now present our main result for this section, which shows that $\hat{\boldsymbol{\psi}}$ is consistent, derives its asymptotic distribution, and establishes consistency of standard errors. Let

$$\mathbf{V} = \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]^{-1} \mathbb{E}\left[\varepsilon_i^2 \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right] \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]^{-1}$$

denote the asymptotic variance of the OLS estimator in the infeasible regression of Y_i on the true latent $\boldsymbol{\theta}_i$ and $\mathbf{q}_i$. Also let

$$\hat{\mathbf{V}} = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T\right) \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T\right)^{-1} \tag{9}$$

denote the covariance matrix estimator for the regression of Y_i on $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{q}_i$.

Theorem 1. *Suppose that Assumption 1 holds. Then as $n \to \infty$:*

1. *The OLS estimator $\hat{\boldsymbol{\psi}}$ from regressing Y_i on $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{q}_i$ is consistent and asymptotically normally distributed, but with a centering that may differ from zero:*

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

2. *Two-step standard errors are consistent:*

$$\hat{\mathbf{V}} \to_p \mathbf{V}. \tag{11}$$

Theorem 1 shows that two-step inference is *invalid* when $\kappa > 0$. In this case, standard errors are consistent but the asymptotic distribution is centered 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.[20] The bias—and thus the degree of under-coverage—increases in κ. Simulations reported in Section 6 show that coverage distortions can be severe even for small values of κ.

[20] 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.

Moreover, because measurement error can be nonclassical, it can be difficult to know even the sign of the bias: there may be attenuation or amplification. These critiques apply to inference on $\boldsymbol{\alpha}$ as well as $\boldsymbol{\gamma}$, and are therefore relevant for researchers using unstructured data to create control variables.

On the other hand, two-step inference is *valid* when $\kappa = 0$. In this case, measurement error is of smaller order than sampling error and can effectively be ignored. Here $\hat{\boldsymbol{\psi}}$ has the same $N(\mathbf{0}, \mathbf{V})$ 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.

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 finite-dimensional parameter that is estimated in the first stage. This across-observation dependence causes the term in (8) to converge to a random variable rather than a constant, leading to the variance inflation.

Remark 2. Our asymptotic framework is related to an econometrics literature on "small" classical measurement error (e.g., Chesher (1991)), in which 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. Their approach imposes no structure on the source of measurement error and uses instrumental variables to identify its variance. Our setting is different: measurement error arises due to first-stage estimation of $\boldsymbol{\theta}_i$, allowing us to analytically characterize and correct bias without an instrument.

We now derive expressions for κ and Ω in the running examples. We shall use these in the next section to perform bias corrections.

4.1.1 Application 1: AI/ML-Generated Labels.

To mimic scenarios where high-performance classifiers are deployed at scale, we consider a sequence of DGPs where the distribution of $(Y_i, \boldsymbol{\theta}_i, \mathbf{q}_i)$ is fixed but the distribution of $\mathbf{x}_i|(Y_i, \boldsymbol{\theta}_i, \mathbf{q}_i)$ varies with n so that $\mathbf{x}_i$ becomes increasingly more informative about $\boldsymbol{\theta}_i$. For brevity we present conditions for the case of a deterministic classifier: $\hat{\theta}_i = \pi(\mathbf{x}_i)$ for some function π taking values in $\{0, 1\}$. We treat π as deterministic, conditioning on the external training data set where necessary. Appendix B.2 presents a more general treatment allowing stochastic classifiers and multiple categories.

Assumption 2. *(i)* $\sqrt{n}\,\mathbb{E}[\hat{\theta}_i(1-\theta_i)] \to \kappa$.
(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) $\mathbb{E}[(\hat{\theta}_i - \theta_i)\mathbf{q}_i] = \mathbf{0}$.

(iv) $\mathbb{E}[\hat{\theta}_i \varepsilon_i] = \mathbf{0}$.

Assumption 2(i) is the key drifting-sequence condition. It says that the false-positive rate $\mathbb{E}[\hat{\theta}_i(1-\theta_i)]$ goes to zero at rate $n^{-1/2}$ (or faster). This allows for the classifier to produce misclassifications which individually occur with low probability, but whose cumulative effect will be non-negligible relative to sampling error when $\kappa > 0$. Assumption 2(ii) is standard. Assumption 2(iii) says $\mathbf{q}_i$ and the prediction errors $\hat{\theta}_i - \theta_i$ are orthogonal. It is straightforward to relax this condition; doing so will simply add nonzero off-diagonal terms in the bias expression in (10) without altering our main point: the two-step strategy can lead to biased inference. Finally, Assumption 2(iv) says the true regression errors ε_i are uncorrelated with the AI/ML-generated prediction $\hat{\theta}_i$. As ε_i and θ_i are assumed uncorrelated, in effect this condition simply requires that the prediction error $\hat{\theta}_i - \theta_i$ and ε_i are uncorrelated.

Theorem 2. *Suppose that Assumption 2 holds. Then Assumption 1 holds and the OLS estimator* $\hat{\boldsymbol{\psi}}$ *has asymptotic distribution given by (10) with* $\kappa = \lim_{n\to\infty} \sqrt{n}\,\mathbb{E}[\hat{\theta}_i(1-\theta_i)]$ *and* $\Omega = 1$. *Moreover, two-step standard errors are consistent.*

4.1.2 Application 2: Topic Models.

In many modern empirical settings, there may be a large number of observations (large n) and a large amount of unstructured data per observation (large C_i). To mimic such scenarios, we consider a sequence of populations in which the distribution of $(Y_i, \mathbf{q}_i, \boldsymbol{w}_i)$ is fixed and the conditional distribution of $(\mathbf{x}_i, C_i)$ given $(Y_i, \mathbf{q}_i, \boldsymbol{w}_i)$ varies with n so that (6) holds and

$$\sqrt{n}\,\mathbb{E}\left[\frac{1}{C_i}\right] \to \kappa. \tag{12}$$

Smaller values of κ correspond to settings where there is a relatively more unstructured data per observation, and hence relatively less measurement error. Also note that since C_i enters via its inverse, if most documents are large but a few are small, then κ may still be large.

To simplify some expressions, in what follows we implicitly 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 sampling error and regression errors are uncorrelated. These assumptions seem very reasonable and can be relaxed: doing so simply complicates the expressions below. We also slightly strengthen (5) to require that $\mathbb{E}[\varepsilon_i(\boldsymbol{w}_i, \mathbf{q}_i)] = \mathbf{0}$. That is, no relevant topic weights have been omitted from the regression.

Assumption 3. *(i)* $\sqrt{n}\,\mathbb{E}\left[\frac{1}{C_i}\right] \to \kappa$.

(ii) $\mathbf{B}$ *has full rank.*
(iii) $\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B}) \to_p \mathbf{0}$.
(iv) $\sqrt{n}\max_{1\le i\le n} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}(\mathbf{x}_i/C_i)\| \to_p 0$.
(v) $\mathbb{E}\left[\|\mathbf{q}_i\|^4\right] < \infty$, $\mathbb{E}\left[\varepsilon_i^4\right] < \infty$, *and* $\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]$ *has full rank.*
(vi) $C_i \gtrsim (\log n)^{1+\epsilon}$ *almost surely for some* $\epsilon > 0$.

Assumption 3(i) is the key drifting sequence condition. Assumption 3(ii) says that none of the topics are redundant. Bing et al. (2020), Wu et al. (2023), and Ke and Wang (2022) show various estimators $\hat{\mathbf{B}}$ converge at the optimal rate $(nC)^{-1/2}$ (up to log terms) where, for simplicity, all C_i are of the same order C (i.e., $C_i \asymp C$). Hence, their estimators all satisfy Assumption 3(iii) when C grows with n, as we have here by (12). Assumption 3(iv) imposes some structure on the $\hat{\boldsymbol{\theta}}_i$ to facilitate derivations. This condition is not vacuous: we have $\boldsymbol{\theta}_i = \boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\mathbb{E}[\mathbf{x}_i/C_i|C_i, \boldsymbol{w}_i]$ by display (6) and Assumption 3(ii). Hence, one could take $\hat{\boldsymbol{\theta}}_i = \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}(\mathbf{x}_i/C_i)$, in which case Assumption 3(iv) trivially holds. Assumption 3(v) is standard. Assumption 3(vi) is made to simplify arguments establishing consistency of standard errors and can be relaxed. This condition trivially holds in view of (12) when all C_i are of the same order, since in that case $C_i \gtrsim n^{1/2}$. We note that these conditions implicitly assume $\mathbf{B}$ is identified. We defer discussion of identification to Appendix B.3.

Theorem 3. *Suppose that Assumption 3 holds. Then Assumption 1 holds and the OLS estimator* $\hat{\boldsymbol{\psi}}$ *has asymptotic distribution given by (10) with* $\kappa = \lim_{n\to\infty}\sqrt{n}\,\mathbb{E}[C_i^{-1}]$ *and*

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

Moreover, two-step standard errors are consistent.

Remark 3. A conceptually related problem involves factor-augmented regressions. In their 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. Bai and Ng (2006) show this approach leads to valid inference provided $\sqrt{T}/N \to 0$, where T is the time-series dimension and N is the cross-sectional dimension.[21] 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, analogous to $\kappa > 0$, then there is a bias that shifts the location of the asymptotic distribution. At an abstract level, Theorem 1 can be seen as generalizing this finding to a broad class of scenarios. Our theory for AI/ML-generated labels and topic models is new and does not follow from these existing works.

[21] See Cahan et al. (2023) and references therein for the related problem of using factor models to impute missing observations.

5 How To Debias Inference

We now propose two methods for performing valid inference on $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$: (1) bias corrected estimators and confidence intervals, and (2) joint estimation of the upstream and downstream regression models. These are the approaches illustrated in the applications in Section 3.

Each method has its strengths and weaknesses. The bias correction is simple to implement and scales well, but the formulas for the bias we develop are for particular settings. While these encompass leading applications, they are not exhaustive, and other formulas will need to be derived for other settings by specializing our general characterization of bias. Joint estimation is more flexible and can handle cases where the general bias formula may be difficult to specialize, but it is also more computationally demanding.

5.1 Bias-Corrected Estimators and Confidence Intervals

Theorem 1 shows that the asymptotic bias of the two-step estimator $\hat{\boldsymbol{\psi}}$ takes the form

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

We use this formula to construct bias-corrected estimators and confidence intervals (CIs) for $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$. Given consistent estimators $\hat{\kappa}$ and $\hat{\boldsymbol{\Omega}}$ of κ and $\boldsymbol{\Omega}$, one can construct the following bias-corrected estimators

$$\hat{\boldsymbol{\psi}}^{bca} = \left(\mathbf{I} + \frac{\hat{\kappa}}{\sqrt{n}} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T \right)^{-1} \begin{bmatrix} \hat{\boldsymbol{\Omega}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right) \hat{\boldsymbol{\psi}},$$

$$\hat{\boldsymbol{\psi}}^{bcm} = \left(\mathbf{I} - \frac{\hat{\kappa}}{\sqrt{n}} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T \right)^{-1} \begin{bmatrix} \hat{\boldsymbol{\Omega}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right)^{-1} \hat{\boldsymbol{\psi}}.$$

The first estimator $\hat{\boldsymbol{\psi}}^{bca}$ performs an additive bias correction to the OLS estimator $\hat{\boldsymbol{\psi}}$. Simulations below show that this estimator performs well when the bias in $\hat{\boldsymbol{\psi}}$ is relatively small. The additive correction may not be sufficient when the bias in $\hat{\boldsymbol{\psi}}$ is large, as it relies upon $\hat{\boldsymbol{\psi}}$ being a reasonable estimator of $\boldsymbol{\psi}$. The bias corrected estimator $\hat{\boldsymbol{\psi}}^{bcm}$ performs a more aggressive (multiplicative) bias correction. We recommend this second estimator when the bias in $\hat{\boldsymbol{\psi}}$ is expected to be large, provided the maximum eigenvalue of

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

is less than one in absolute value.[22] We discuss how to consistently estimate κ and $\mathbf{\Omega}$ within the context of our running examples below.

Bias corrected CIs for the regression coefficients are constructed by centering at the bias-corrected estimators and using OLS standard errors. A valid $100(1-a)\%$ CI for the jth component ψ_j of $\boldsymbol{\psi}$ is given by

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

where $\hat{\psi}_j^{bc}$ denotes the jth entry of $\hat{\boldsymbol{\psi}}^{bca}$ or $\hat{\boldsymbol{\psi}}^{bcm}$, $z_{1-a/2}$ is the $1-a/2$ 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 $\hat{\mathbf{V}}$ from (9).

The following result shows that the bias-corrected estimators are asymptotically normal with the correct centering, and the bias-corrected confidence intervals are valid:

Theorem 4 (Validity of Bias-Corrected Inference)**.** *Suppose that Assumption 1 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* $\hat{\mathbf{\Omega}} \to_p \mathbf{\Omega}$*. Then as* $n \to \infty$*:*

1. *Bias-corrected estimators are first-order asymptotically equivalent and asymptotically normally distributed with the correct centering:*

$$\sqrt{n}\left(\hat{\boldsymbol{\psi}}^{bcm} - \boldsymbol{\psi}\right) = \sqrt{n}\left(\hat{\boldsymbol{\psi}}^{bca} - \boldsymbol{\psi}\right) + o_p(1) \to_d N(\mathbf{0}, \mathbf{V});$$

2. *Bias-corrected confidence intervals have correct coverage:*

$$\lim_{n\to\infty} \Pr(\psi_j \in \mathrm{CI}_n(\psi_j)) = 1 - a.$$

We now show how to apply bias correction in the context of our two running examples. In each case, we propose consistent estimators $\hat{\kappa}$ and $\hat{\mathbf{\Omega}}$ of κ and $\mathbf{\Omega}$. Theorem 4 then implies that the bias-corrected CIs have correct coverage. We also recommend reporting $\hat{\kappa}$ as a diagnostic for assessing the relative importance of measurement error and sampling error.

5.1.1 Application 1: AI/ML-Generated Labels.

Here $\kappa = \lim_{n\to\infty}\sqrt{n}\,\mathbb{E}[\hat{\theta}_i(1-\theta_i)]$ and $\mathbf{\Omega} = 1$. One may estimate κ by taking a sample of observations of size $m \ll n$. Then, each observation $i = 1, \ldots, m$ for which $\hat{\theta}_i = 1$ is

[22] This condition ensures invertibility of $\left(\mathbf{I} - \frac{\hat{\kappa}}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1}\begin{bmatrix}\hat{\mathbf{\Omega}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right)$.

inspected, and assigned a true label θ_i. The estimator of κ is

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

We establish validity of this approach allowing $m/n \to 0$ asymptotically. This is important for accommodating modern use cases where ML/AI methods are deployed to impute labels on massive data sets (large n), but where correctly labeling data can be costly (small m). 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 (e.g., Fong and Tyler (2021), Allon et al. (2023), Egami et al. (2023)) are shown to be valid when $m/n \to c > 0$, so that the validation and original sample sizes are comparable. As far as we are aware, the theoretical properties of these proposed methods are unknown in modern scenarios where $m/n \to 0$.

We emphasize that our approach does not require constructing a full validation sample of size m (and thus performing costly inspection of all m observations), but only those for which $\hat{\theta}_i = 1$. This can greatly reduce the burden on the researcher. For instance, in our empirical application to remote work, we take a subsample of size $m = 1000$. Of these, only 26 observations have $\hat{\theta}_i = 1$, so only 26 job postings need to be inspected. By contrast, constructing a validation sample would require inspecting all $m = 1000$ observations.

Finally, our bias correction can be implemented using external data in which Y_i and/or $\mathbf{q}_i$ are missing, since it only requires a subsample of θ_i and $\hat{\theta}_i$. This makes our approach more broadly applicable than other methods that require a full validation data set. For instance, Bursztyn et al. (2024) use a ML algorithm to classify a data set of charitable donors' names by ethnicity. As true ethnicity is latent, they estimate the accuracy of the classifier using an external sample of North Carolina voter registration data which contains self-reported ethnicity (but is missing data on charitable donations).

The following result shows that $\hat{\kappa}$ is consistent, allowing $m/n \to 0$.

Lemma 1. *Suppose that* $\sqrt{n}\,\mathbb{E}[\hat{\theta}_i(1-\theta_i)] \to \kappa > 0$ *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 $\hat{\kappa}$ from a small subsample can introduce additional variability that, while asymptotically negligible, may be important to account for in finite samples. To this end, we introduce the following finite-sample correction to standard errors. Recall $\hat{\mathbf{V}}$ from (9).

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 estimators $\hat{\mathbf{V}}^{bca}$ and $\hat{\mathbf{V}}^{bcm}$ for $\hat{\boldsymbol{\psi}}^{bca}$ and $\hat{\boldsymbol{\psi}}^{bcm}$ are given by

$$\hat{\mathbf{V}}^{bca} = (\mathbf{I} + \widehat{FPR}\,\hat{\mathbf{\Gamma}})\hat{\mathbf{V}}(\mathbf{I} + \widehat{FPR}\,\hat{\mathbf{\Gamma}}^T) + \frac{1}{m}\widehat{FPR}(1-\widehat{FPR})\hat{\mathbf{\Gamma}}\left(\hat{\mathbf{V}} + n\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^T\right)\hat{\mathbf{\Gamma}}^T,$$
$$\hat{\mathbf{V}}^{bcm} = (\mathbf{I} - \widehat{FPR}\,\hat{\mathbf{\Gamma}})^{-1}\hat{\mathbf{V}}(\mathbf{I} - \widehat{FPR}\,\hat{\mathbf{\Gamma}}^T)^{-1} + \frac{1}{m}\widehat{FPR}(1-\widehat{FPR})\hat{\mathbf{\Gamma}}\left(\hat{\mathbf{V}} + n\hat{\boldsymbol{\psi}}\hat{\boldsymbol{\psi}}^T\right)\hat{\mathbf{\Gamma}}^T,$$

respectively. The form of these adjustments follows from the law of total variance. Under the conditions of Theorem 1 and Lemma 1, we have $\hat{\mathbf{V}}^{bca} \to_p \mathbf{V}$ and $\hat{\mathbf{V}}^{bcm} \to_p \mathbf{V}$, with $\mathbf{V}$ the asymptotic variance derived in Theorem 1. In practice, we recommend reporting standard errors computed from $\hat{\mathbf{V}}^{bca}$ if using $\hat{\boldsymbol{\psi}}^{bca}$ or $\hat{\mathbf{V}}^{bcm}$ if using $\hat{\boldsymbol{\psi}}^{bcm}$.

Codes to implement these bias corrections and standard error formulas are available in the Python package `ValidMLInference`.

5.1.2 Application 2: Topic Models.

In view of the expressions for κ and $\mathbf{\Omega}$ derived in Theorem 3, consider

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

where $\bar{\boldsymbol{w}}_n = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{w}}_i$. The following result shows that these estimators are consistent.

Lemma 2. *Suppose that Assumption 3 holds and that* $\bar{\boldsymbol{w}}_n \to_p \mathbb{E}[\boldsymbol{w}_i]$. *Then* $\hat{\kappa} \to_p \kappa$ *and* $\hat{\mathbf{\Omega}} \to_p \mathbf{\Omega}$.

5.2 Joint Estimation

Our second approach for correcting the bias from the two-step strategy begins by formulating a likelihood $l\left(Y_i, h(\mathbf{x}_i), \boldsymbol{\theta}_i \mid \mathbf{q}_i, \mathbf{v}_i, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\zeta}\right)$. Here h is a function of the high-dimensional or unstructured observables, $\mathbf{v}_i$ are covariates that potentially enter the model beyond the downstream regression, and $\boldsymbol{\zeta}$ are nuisance parameters. We discuss how to form a likelihood below. Integrating the latent $\boldsymbol{\theta}_i$ out yields a likelihood $l\left(Y_i, h(\mathbf{x}_i) \mid \mathbf{q}_i, \mathbf{v}_i, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\zeta}\right)$ depending only on observables, which can then be used for maximum likelihood estimation of model parameters. This idea can be applied generically, and here we adapt it to each of the three applications in Section 3.

5.2.1 Remote Work and Wage Inequality

In the remote work application, Y_i is log posted wages, $\theta_i \in \{0, 1\}$ is an indicator for whether job posting i offers remote work ($\theta_i = 1$) or not ($\theta_i = 0$), and $\mathbf{x}_i$ is job posting text. We use $h(\mathbf{x}_i) = \hat{\theta}_i \in \{0, 1\}$ and thereby formulate a likelihood over the predicted class label associated with $\mathbf{x}_i$ rather than over $\mathbf{x}_i$ directly. No additional covariates enter the model so $\mathbf{v}_i$ is empty. The integrated likelihood is

$$l(Y_i, \hat{\theta}_i = d; (\gamma, \boldsymbol{\alpha}, \boldsymbol{\zeta})) = \omega_{d1} MN(Y_i - \gamma - \boldsymbol{\alpha}^T \mathbf{q}_i; \boldsymbol{\lambda}_1) + \omega_{d0} MN(Y_i - \boldsymbol{\alpha}^T \mathbf{q}_i; \boldsymbol{\lambda}_0), \tag{13}$$

for $d \in \{0, 1\}$, $\boldsymbol{\omega} = (\omega_{00}, \omega_{10}, \omega_{01}, \omega_{11}) \in \Delta^3$ with $\omega_{ab} = \Pr(\hat{\theta}_i = a, \theta_i = b)$, and $MN(\,\cdot\,; \boldsymbol{\lambda}) = \sum_{l=1}^{L} \lambda_l \phi(\,\cdot\,; \lambda_{\mu l}, \lambda_{\sigma^2 l})$ is a mixture of normal distributions with L components and mixing weights $\lambda_1, \ldots, \lambda_L \in \Delta^{L-1}$, where $\phi(\,\cdot\,; \mu, \sigma^2)$ denotes the $N(\mu, \sigma^2)$ density, and the component means are normalized so that $\sum_{l=1}^{L} \lambda_l \lambda_{\mu l} = 0$. Thus, here $\boldsymbol{\zeta} = (\boldsymbol{\omega}, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1)$. We use $L = 3$ for the results in Section 3.1.

5.2.2 CEO Time Use and Firm Performance

In the CEO time use application, Y_i is log sales of firm i, $\theta_i \in [0, 1]$ is a behavior index, and $\mathbf{x}_i$ is the vector of counts of time-use feature combinations across all surveyed 15-minute time intervals. Here the likelihood is directly over $\mathbf{x}_i$, i.e. $h(\mathbf{x}_i) = \mathbf{x}_i$. In the spirit of correlated random effects in panel data models, we specify a distribution for the behavior index θ_i conditional on J covariates $\mathbf{g}_i$, which may include $\mathbf{q}_i$ or other variables.[23] The likelihood is implied by

$$\begin{aligned} \theta_i \mid (C_i, \mathbf{q}_i, \mathbf{g}_i) &\sim \text{LogisticNormal}\left(\boldsymbol{\phi}^T \mathbf{g}_i, \sigma_\theta^2\right), \\ \mathbf{x}_i \mid (\theta_i, C_i, \mathbf{q}_i, \mathbf{g}_i) &\sim \text{Multinomial}\left(C_i, \mathbf{B}^T \mathbf{w}_i\right), \\ Y_i \mid (\theta_i, C_i, \mathbf{q}_i, \mathbf{g}_i) &\sim \text{Normal}\left(\gamma\theta_i + \boldsymbol{\alpha}^T \mathbf{q}_i, \sigma_Y^2\right), \end{aligned} \tag{14}$$

where $\mathbf{w}_i = (1 - \theta_i, \theta_i)^T$. Bandiera et al. (2020) shows log employment and an indicator for whether the CEO has an MBA are correlated with behavior, so we include these in $\mathbf{g}_i$ together with a constant. As the likelihood also depends on the number of observed time units C_i, we have $\mathbf{v}_i = (C_i, \mathbf{g}_i)$.

[23] Roberts et al. (2014) presents a model in which a logistic normal distribution over topic shares 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 topic shares.

5.2.3 Central Bank Communication

In this application, Y_i is the path factor in FOMC meeting i, $\theta_i \in [0,1]$ is the hawkish sentiment of the FOMC written statement, and $\mathbf{x}_i$ is the text of the statement. We take $h(\mathbf{x}_i) = (N_i, C_i)$ and form a likelihood over the number of paragraphs in FOMC statements classified as hawkish (N_i) conditional on the total number classified as hawkish or dovish (C_i). We treat $\mathbf{v}_i$ as empty, but our analysis could be extended to allow the prior for θ_i to depend on covariates. The likelihood is implied by

$$\begin{aligned} \theta_i \mid (C_i, \mathbf{q}_i) &\sim U[0,1], \\ N_i \mid (\theta_i, C_i, \mathbf{q}_i) &\sim \text{Binomial}\left(C_i, (1-\theta_i)\beta_0 + \theta_i\beta_1\right), \\ Y_i \mid (\theta_i, C_i, \mathbf{q}_i) &\sim \text{Normal}\left(\gamma\theta_i + \boldsymbol{\alpha}^T\mathbf{q}_i, \sigma_Y^2\right). \end{aligned} \tag{15}$$

In addition, we include additional terms in the likelihood for the testing data:

$$N_{10} \sim \text{Binomial}(N_{10}+N_{00}, \beta_0), \quad N_{11} \sim \text{Binomial}(N_{11}+N_{01}, \beta_1), \tag{16}$$

where N_{ab} is the number of test-set observations classified an a with true label b.[24]

5.3 Inference Approach for Intractable Likelihoods

In certain cases, one can directly integrate-out $\boldsymbol{\theta}_i$ from the likelihood, as in (13). However, in more complex problems, such as (14), there are two challenges. First, the integration for computing the likelihood has no closed-form solution and must be performed numerically. Moreover, this numerical integration must be done observation-by-observation. As such, standard likelihood-based estimation may not be computationally feasible.

In such cases, one may use Bayesian computation to perform valid frequentist inference. In this approach, we introduce a prior for the model parameters $\boldsymbol{\delta} = (\boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\zeta})$ and treat the latent $\boldsymbol{\theta}_i$ as "parameters" drawn from a prior distribution as illustrated above. We sample from the posterior distribution of $(\boldsymbol{\delta}, (\boldsymbol{\theta}_i)_{i=1}^n)$ conditional on the observed data $(Y_i, h(\mathbf{x}_i), \mathbf{q}_i, \mathbf{v}_i)_{i=1}^n$. The marginal draws for $\boldsymbol{\delta}$ represent draws from the posterior distribution for $\boldsymbol{\delta}$ based on the integrated likelihood $l\left(Y_i, h(\mathbf{x}_i) \mid \mathbf{q}_i, \mathbf{v}_i, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\zeta}\right)$. Thus, $\boldsymbol{\theta}_i$ is implicitly integrated of the likelihood as part of the sampling procedure.

It is important to emphasize that while our suggested approach uses Bayesian computation, inference is frequentist. 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

[24] In the application, $N_{11} = 14$, $N_{00} = 54$, $N_{10} = 1$, $N_{01} = 1$.

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 elements of $\boldsymbol{\delta}$—or any of its components such as $\boldsymbol{\gamma}$ or $\boldsymbol{\alpha}$—are valid frequentist confidence sets. This approach is also *efficient* for inference on $\boldsymbol{\delta}$ and its components, as it is asymptotically equivalent to likelihood-based inference.

We implement this approach for the CEO time use and central bank communication applications. For the former, we follow standard practice in the topic modeling literature and use Dirichlet(0.2) priors for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. For the logistic normal prior for $\boldsymbol{\theta}_i$, we set $\sigma_\theta^2 = 1$ and use a Normal$(0, 4)$ prior for each element of $\boldsymbol{\phi}$. We also use Normal$(0, 100)$ priors for all downstream regression coefficients, and a Gamma$(1, 10)$ prior for σ_Y.

For the central bank communication application, we use the same priors in the downstream regression. We use a Beta$(2, 5)$ prior for β_0 and a Beta$(5, 2)$ prior for β_1. We adopt asymmetric priors on the classification probabilities to resolve the label switching problem inherent in topic models.

We perform sampling with Hamiltonian Monte Carlo implemented in `NumPyro` (Phan et al. 2019). See Sacher et al. (2024) for more details on HMC and probabilistic programming.

6 Simulation Evidence

In this section, we provide simulation evidence illustrating the finite-sample performance of our methods for correcting bias and performing valid inference.

6.1 AI/ML-Generated Labels

This simulation is calibrated to the remote work empirical application in Section 3.1. The same contains 16,315 job postings. Bias-corrected estimates of the intercept and slope are approximately 10 and 1. The residual standard deviation is approximately 0.3 (respectively, 0.5) for observations with $\hat{\theta}_i = 0$ ($\hat{\theta}_i = 1$). We therefore generate data according to

$$Y_i = 10 + \theta_i + (0.3 + 0.2\theta_i)\varepsilon_i,$$

with $\varepsilon_i \sim N(0, 1)$. The sample mean of $\hat{\theta}_i$ is 0.025 and $\widehat{FPR} = 0.009$, which corresponds to $\hat{\kappa} \approx 1.1$. We draw samples of size $n =$ 8,000, 16,000, and 32,000, with $\theta_i \sim$ Bernoulli(p) for $p = 0.025$, 0.05, and 0.5, and $\kappa = 0.5$, 1, and 2, with 1000 simulations for each configuration.

Table 4: Simulation Results: Generated Labels, $p = 0.025$

	Bias			RMdSE			Coverage		
κ	0.5	1	2	0.5	1	2	0.5	1	2
				$n = 8000$					
2-Step	-0.228	-0.459	-0.916	0.228	0.459	0.916	0.001	0.000	0.000
BCA-1	-0.055	-0.214	-0.843	0.076	0.214	0.843	0.899	0.589	0.000
BCA-2	-0.039	-0.203	-0.841	0.073	0.204	0.841	0.929	0.643	0.000
BCM-1	-0.003	-0.006	-0.715	0.091	0.170	0.806	0.887	0.758	0.265
BCM-2	0.023	0.030	-0.729	0.090	0.177	0.827	0.904	0.773	0.257
Joint	0.000	0.002	-0.008	0.045	0.056	0.111	0.935	0.930	0.817
				$n = 16000$					
2-Step	-0.161	-0.323	-0.648	0.161	0.323	0.648	0.000	0.000	0.000
BCA-1	-0.028	-0.110	-0.423	0.058	0.111	0.423	0.884	0.794	0.049
BCA-2	-0.011	-0.097	-0.417	0.055	0.101	0.417	0.928	0.841	0.058
BCM-1	-0.001	-0.005	-0.012	0.071	0.110	0.303	0.874	0.832	0.542
BCM-2	0.024	0.024	0.037	0.070	0.114	0.313	0.915	0.851	0.548
Joint	0.002	0.002	0.002	0.030	0.034	0.055	0.952	0.951	0.897
				$n = 32000$					
2-Step	-0.116	-0.231	-0.459	0.116	0.231	0.459	0.000	0.000	0.000
BCA-1	-0.022	-0.056	-0.214	0.048	0.069	0.214	0.878	0.859	0.505
BCA-2	-0.004	-0.040	-0.204	0.044	0.062	0.204	0.938	0.907	0.566
BCM-1	-0.009	-0.007	-0.015	0.055	0.085	0.165	0.874	0.868	0.735
BCM-2	0.014	0.019	0.022	0.054	0.085	0.168	0.927	0.896	0.758
Joint	-0.002	0.000	0.001	0.019	0.024	0.027	0.950	0.952	0.957

Note: Median bias (Bias), root median square error (RMdSE), and coverage of 95% confidence intervals (Coverage) across different κ and n. Results are presented for the two-step strategy (2-step), additive bias correction using $\widehat{FPR}$ (BCA-1) and $\widehat{FPR}_B$ (BCA-2), multiplicative bias correction using $\widehat{FPR}$ (BCM-1) and $\widehat{FPR}_B$ (BCM-2), and joint estimation (Joint).

We implement the two-step strategy, additive and multiplicative bias corrections, and joint estimation. For the latter, we use the likelihood from (13) with a single Gaussian component ($L = 1$). To implement both bias corrections, we generate a sample of $(\theta_i, \hat{\theta}_i)$ of size $m = 1000$. We use two estimators of the false-positive rate. The first is the empirical frequency $\widehat{FPR} = \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i(1-\theta_i)$. We may interpret this as the posterior mean of $r := \mathbb{E}[\hat{\theta}_i(1-\theta_i)]$ under an improper $r^{-1}(1-r)^{-1}$ prior. This prior puts most of its mass at the endpoints of the interval $[0, 1]$. However, our approach is based on the premise that r should be small. We therefore consider a Bayes estimator $\widehat{FPR}_B = \frac{\sum_{i=1}^{m} \hat{\theta}_i(1-\theta_i)+\frac{1}{2}}{m+\frac{5}{2}}$, which is the posterior mean of r under a proper $r^{-1/2}(1-r)$ prior. Results are presented in Tables 4-6.

Table 5: Simulation Results: Generated Labels, $p = 0.05$

	Bias			RMdSE			Coverage		
κ	0.5	1	2	0.5	1	2	0.5	1	2
				$n = 8000$					
2-Step	-0.116	-0.236	-0.469	0.116	0.236	0.469	0.019	0.000	0.000
BCA-1	-0.015	-0.058	-0.224	0.039	0.065	0.224	0.931	0.868	0.319
BCA-2	-0.005	-0.050	-0.219	0.037	0.060	0.219	0.956	0.893	0.335
BCM-1	-0.002	-0.005	-0.010	0.043	0.063	0.131	0.927	0.884	0.751
BCM-2	0.010	0.008	0.009	0.041	0.064	0.130	0.944	0.899	0.760
Joint	0.000	-0.003	-0.004	0.028	0.031	0.043	0.956	0.951	0.927
				$n = 16000$					
2-Step	-0.081	-0.165	-0.332	0.081	0.165	0.332	0.018	0.000	0.000
BCA-1	-0.007	-0.031	-0.110	0.031	0.044	0.110	0.916	0.893	0.690
BCA-2	0.003	-0.022	-0.103	0.029	0.040	0.104	0.948	0.917	0.719
BCM-1	-0.001	-0.004	0.000	0.033	0.048	0.090	0.910	0.886	0.817
BCM-2	0.011	0.008	0.014	0.034	0.049	0.088	0.938	0.915	0.823
Joint	0.000	0.001	0.002	0.018	0.021	0.023	0.948	0.953	0.935
				$n = 32000$					
2-Step	-0.059	-0.118	-0.236	0.059	0.118	0.236	0.006	0.000	0.000
BCA-1	-0.009	-0.015	-0.060	0.025	0.034	0.063	0.897	0.904	0.835
BCA-2	0.001	-0.006	-0.052	0.024	0.032	0.056	0.949	0.935	0.864
BCM-1	-0.006	-0.004	-0.005	0.027	0.041	0.065	0.892	0.902	0.871
BCM-2	0.005	0.008	0.008	0.027	0.039	0.062	0.938	0.929	0.880
Joint	-0.001	-0.001	0.000	0.013	0.015	0.016	0.957	0.949	0.961

Note: Median bias (Bias), root median square error (RMdSE), and coverage of 95% confidence intervals (Coverage) across different κ and n. Results are presented for the two-step strategy (2-step), additive bias correction using $\widehat{FPR}$ (BCA-1) and $\widehat{FPR}_B$ (BCA-2), multiplicative bias correction using $\widehat{FPR}$ (BCM-1) and $\widehat{FPR}_B$ (BCM-2), and joint estimation (Joint).

Our baseline set of simulations with $p = 0.025$ (Table 4) is very challenging: in a sample of size 8,000, we expect only 200 observations with $\theta_i = 1$. When $\kappa = 1$ about half of these will be incorrectly imputed with $\hat{\theta}_i = 0$, and when $\kappa = 2$ almost all will be incorrectly imputed. In either case, we expect a very large bias of two-step estimators, and the results in Table 4 confirm this, with large bias and zero coverage. Our bias corrections rely on the two-step estimator having a bias that is of the same order as sampling uncertainty. One might therefore expect they will perform poorly in small samples when two-step estimators are severely biased. Table 4 shows this is not necessarily the case. Consider the configuration with $n = 16{,}000$ and $\kappa = 1$, which is closest to the empirical application. The two-step

Table 6: Simulation Results: Generated Labels, $p = 0.5$

	Bias			RMdSE			Coverage		
κ	0.5	1	2	0.5	1	2	0.5	1	2
				$n = 8000$					
2-Step	-0.022	-0.045	-0.089	0.022	0.045	0.089	0.351	0.002	0.000
BCA-1	0.000	-0.002	-0.008	0.009	0.011	0.014	0.944	0.943	0.925
BCA-2	0.002	0.000	-0.007	0.009	0.010	0.014	0.957	0.956	0.937
BCM-1	0.000	0.000	0.000	0.010	0.011	0.016	0.942	0.941	0.925
BCM-2	0.002	0.002	0.002	0.009	0.012	0.016	0.954	0.949	0.932
Joint	0.000	0.000	0.000	0.008	0.008	0.009	0.944	0.956	0.954
				$n = 16000$					
2-Step	-0.015	-0.032	-0.063	0.015	0.032	0.063	0.344	0.002	0.000
BCA-1	0.000	-0.001	-0.004	0.007	0.009	0.012	0.933	0.950	0.930
BCA-2	0.002	0.000	-0.003	0.007	0.009	0.012	0.956	0.955	0.950
BCM-1	0.000	0.000	0.000	0.007	0.009	0.013	0.931	0.944	0.935
BCM-2	0.002	0.002	0.002	0.007	0.009	0.012	0.950	0.949	0.945
Joint	0.000	0.000	0.000	0.005	0.005	0.006	0.947	0.944	0.960
				$n = 32000$					
2-Step	-0.011	-0.023	-0.045	0.011	0.023	0.045	0.322	0.005	0.000
BCA-1	-0.001	-0.001	-0.002	0.005	0.007	0.009	0.919	0.920	0.927
BCA-2	0.001	0.001	0.000	0.005	0.007	0.009	0.955	0.948	0.938
BCM-1	-0.001	0.000	0.000	0.005	0.008	0.010	0.917	0.918	0.921
BCM-2	0.001	0.002	0.002	0.005	0.007	0.010	0.951	0.943	0.931
Joint	0.000	0.000	0.000	0.004	0.005	0.004	0.943	0.932	0.943

Note: Median bias (Bias), root median square error (RMdSE), and coverage of 95% confidence intervals (Coverage) across different κ and n. Results are presented for the two-step strategy (2-step), additive bias correction using $\widehat{FPR}$ (BCA-1) and $\widehat{FPR}_B$ (BCA-2), multiplicative bias correction using $\widehat{FPR}$ (BCM-1) and $\widehat{FPR}_B$ (BCM-2), and joint estimation (Joint).

estimator under-estimates the coefficient by about 32%, the additive bias corrections by about 10%, while the multiplicative bias corrected estimators are approximately unbiased, even in this challenging design. Indeed, the multiplicative bias correction appears to perform well across all configurations except when n is smallest and κ is largest.

Bias corrected confidence sets appear to under-cover by about 10% with $n = 16{,}000$ and $\kappa = 1$, but the results for $n = 32{,}000$ show their coverage moves toward nominal coverage as n increases. Coverage is much closer to nominal coverage in the less challenging designs with $p = 0.05$ (Table 5) and $p = 0.5$ (Table 6), even with small n. Coverage also seems to be improved using the Bayes estimator $\widehat{FPR}_B$ instead of the empirical frequency.

Finally, joint estimation produces approximately unbiased estimators across all designs, including the most challenging cases. It also produces confidence sets with coverage closest to nominal coverage and, as expected, the lowest risk of the competing approaches. Overall, these results demonstrate the sound performance of both our proposed solutions in a challenging, empirically calibrated setting.

6.2 Topic Models and AI/ML-Generated Indices

These simulations are calibrated to the empirical application to AI/ML-generated indices in Section 3.1. We draw a latent share $\theta_i \sim U[0,1]$ then generate $N_i \sim \text{Binomial}(C_i, \beta_1\theta_i + \beta_0(1-\theta_i))$ with $\beta_1 = 0.9$ and $\beta_0 = 0.1$ to introduce misclassification error. We set

$$Y_i = -0.05 + 0.11\theta_i + 0.1\varepsilon_i,$$

where $\varepsilon_i \sim N(0,1)$ and parameters are similar to the estimates in the empirical application. We generate samples of (Y_i, N_i, C_i) of size $n = 200$ (as in the application), 800, and 3,200. For $n = 200$ we use C_i from the empirical application so that $\kappa \approx 4.57$. We then increase C_i by a factors of two and four, to generate samples with $\kappa \approx 2.28$ and $\kappa \approx 1.14$, respectively. For the larger sample sizes we replicate the empirical C_i and multiply by factors of two and four so κ remains constant. We generate 1000 samples for each configuration.

We implement the two-step strategy, both bias corrections, and joint estimation. For the latter, we use the likelihood described in (15). For the two-step strategy, we require a suitable estimate of θ_i. We produce estimates based on the topic model representation from Section 3 in order to account for potential misclassification error. To provide a clear comparison with joint estimation, we use the estimates of β_0 and β_1 from joint estimation, then estimate θ_i in accordance with the discussion in Section 4.1.2, using

$$\hat{\theta}_i = \max(0, \min(1, \boldsymbol{S}(\hat{\mathbf{B}}^T)^{-1}(\mathbf{n}_i/C_i))),$$

where

$$\boldsymbol{S} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad \hat{\mathbf{B}} = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_0 \\ 1-\hat{\beta}_1 & 1-\hat{\beta}_0 \end{bmatrix}, \quad \mathbf{n}_i = \begin{bmatrix} N_i \\ C_i - N_i \end{bmatrix}.$$

We also implement additive and multiplicative bias corrections using the formulas derived for topic models with the $\boldsymbol{S}$ and $\hat{\mathbf{B}}$ as above. Results are presented in Table 7.

First consider the bias results. In our baseline setting with $n = 200$ and $\kappa = 4.57$ as in the empirical application, the median relative bias of the two-step estimate of γ is -0.446, showing that the two-step strategy under-estimates γ by nearly half. By contrast, joint

Table 7: Simulation Results: AI/ML-Generated Indices

	Bias			RMdSE			Coverage		
κ	4.57	2.28	1.14	4.57	2.28	1.14	4.57	2.28	1.14
				$n = 200$					
2-Step	-0.446	-0.299	-0.180	0.049	0.033	0.022	0.309	0.677	0.839
BCA	-0.148	-0.019	0.018	0.026	0.021	0.019	0.716	0.813	0.880
BCM	0.240	0.183	0.081	0.040	0.029	0.021	0.495	0.678	0.844
Joint	-0.003	0.007	0.004	0.024	0.020	0.018	0.945	0.948	0.938
2-Step-Share	-0.433	-0.218	-0.037	0.048	0.025	0.018	0.378	0.824	0.931
				$n = 800$					
2-Step	-0.301	-0.193	-0.113	0.033	0.021	0.013	0.147	0.533	0.806
BCA	-0.015	0.002	0.005	0.011	0.010	0.010	0.823	0.880	0.911
BCM	0.182	0.063	0.023	0.021	0.011	0.010	0.507	0.798	0.897
Joint	0.004	-0.006	-0.006	0.011	0.010	0.010	0.956	0.950	0.950
2-Step-Share	-0.215	-0.041	0.084	0.024	0.01	0.012	0.507	0.942	0.894
				$n = 3200$					
2-Step	-0.194	-0.110	-0.060	0.021	0.012	0.007	0.053	0.456	0.773
BCA	-0.001	0.007	0.005	0.005	0.005	0.005	0.859	0.896	0.917
BCM	0.060	0.025	0.010	0.007	0.005	0.005	0.700	0.869	0.913
Joint	-0.005	-0.002	-0.003	0.005	0.005	0.005	0.942	0.941	0.943
2-Step-Share	-0.042	0.085	0.158	0.006	0.009	0.017	0.887	0.739	0.353

Note: Median relative bias (Bias), root median square error (RMdSE), and coverage of 95% confidence intervals (Coverage) across different κ and n. Results are presented for the two-step strategy (2-step), additive and multiplicative bias corrections (BCA) and (BCM), joint estimation (Joint), and the two-step strategy regressing onto the empirical share $\hat{\theta}_i = N_i/C_i$ (2-Step-Share).

estimates are nearly unbiased. The additive bias correction improves upon the two-step strategy, producing an estimate about 85% that of the true effect size when $\kappa = 4.57$, and approximately unbiased estimates for smaller κ. Conversely, the multiplicative correction tends to over-estimate the true effect size, suggesting it performs too aggressive a bias correction in small samples. In larger samples, the additive correction produces approximately unbiased estimates, and the performance of the multiplicative correction improves, especially for smaller values of κ. The second panel of Table 7 also suggests the additive bias correction is preferred from a root median square error perspective, producing values that are on par with joint estimation.

Turning to coverage, we see that the two-step CIs have coverage well below nominal coverage, especially for larger values of n and κ. Conversely, joint estimation produces CIs

Table 8: Additional Results: Remote Work Application

	No Fixed Effects			With Fixed Effects		
	Estimate	Std Err	95% CI	Estimate	Std Err	95% CI
BCA-1	0.897	0.119	[0.663, 1.131]	0.521	0.081	[0.362, 0.680]
BCA-2	0.910	0.124	[0.667, 1.154]	0.530	0.084	[0.365, 0.694]
BCM-1	1.052	0.140	[0.778, 1.327]	0.641	0.100	[0.446, 0.836]
BCM-2	1.088	0.148	[0.798, 1.379]	0.668	0.106	[0.460, 0.876]

Note: Results are presented for the additive bias correction using $\widehat{FPR}$ (BCA-1) and $\widehat{FPR}_B$ (BCA-2), and multiplicative bias correction using $\widehat{FPR}$ (BCM-1) and $\widehat{FPR}_B$ (BCM-2).

with coverage close to nominal coverage even for the smallest sample size. Coverage of the additively bias-corrected CIs is significantly better than that of the multiplicatively bias-corrected CIs which, in turn, is better than two-step coverage. For large n and small κ, CIs based on the additive bias correction have coverage close to nominal coverage.

6.3 Lessons for the Empirical Applications

We conclude this section by revisiting some of the empirical applications in light of the simulation evidence presented above. First consider the application to remote work (Section 3.1). The bias-corrected estimates reported in Section 3.1 are based on the multiplicative bias correction, which the above results showed performed best in simulations mimicking this design, and the empirical estimate $\widehat{FPR}$. Results for the different bias corrections with different estimates of the false-positive rate are presented in Table 8. As before, estimated effect sizes are larger for the multiplicative correction, and the multiplicatively corrected CIs are slightly to the right of those using the additive correction. Additively corrected CIs overlap slightly with two-step CIs reported in Section 3.1, while multiplicatively corrected CIs do not.

In view of the simulation evidence above for the topic model, the bias corrections reported in Section 3.2 use the additive correction, but the multiplicative correction produced values that agreed to within ± 0.01.

For central bank communication application in Section 3.3, we implemented the two-step strategy using the empirical share $\hat{\theta}_i = N_i/C_i$ to estimate θ_i. Table 7 presents a set of simulation results for this case as well. This share estimate is prone to two sources of measurement error: misclassification error and upstream sampling error. The results in Table 7 hold the former fixed and let the latter go to zero appropriately with the sample size. Evidently, these measurement errors work in different directions. With small n and large κ,[25]

[25]Note that the κ here is defined within the context of the topic model for the simulations in Section 6.2,

upstream sampling error is more important. This is a "classical" measurement error that causes attenuation bias. With $n = 200$ and $\kappa = 4.57$, as in the empirical application, the bias is comparable to that obtained by implementing the two-step strategy using the topic model, underestimating the true effect size by around 43%. However, bias goes in a different direction when n is large and κ is small, so that misclassification error is more important than upstream sampling error. For instance, with $n = 3{,}200$ and $\kappa = 1.14$, two-step regression onto shares now *over-estimates* the true effect size by around 16%. Correspondingly, the two-step confidence interval has coverage well below nominal coverage.

An important take-away from these results is that the bias of two-step estimators using AI/ML-generated variables can behave differently than in classical measurement error settings, making it difficult for researchers to determine even the sign of the bias.

7 Conclusion

The leading approach for analyzing unstructured or high-dimensional data follows a two-step strategy. 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, how 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 substantial in practice.

To address this problem, we propose two robust alternative inference methods: (1) an explicit bias correction with bias-corrected confidence intervals; and (2) joint maximum likelihood estimation. In a series of simulations and applications involving label imputation, dimensionality reduction, and index construction via classification and aggregation, we show that the two-step strategy produces material biases whereas both proposed methods perform well.

where upstream sampling error is the only source of measurement error. An appropriate κ as in (8) for the share variable $\hat{\theta}_i = N_i/C_i$ must account for both misclassification error and upstream sampling error.

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 C.

Proof of Theorem 1. We first prove part 2. We have $\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T \to_p \mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]$ by Assumption 1(ii). Result (11) now follows by Assumptions 1(i) and 1(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 1(i)-(iii) and the Continuous Mapping Theorem that

$$T_{2,n} \to_d N(\mathbf{0},\mathbf{V}).$$

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 1(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 1(i) and 1(ii) and Slutsky's theorem. ∎

Proof of Theorem 2. This is a special case of Theorem 7. ∎

Proof of Theorem 3. Assumption 1(i) is implied by Assumption 3(v).

The second and third parts of Assumption 1(ii) hold by Lemmas 6 and 7 in Appendix C. 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 $\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T]$ by Assumption 3(v), while the first two terms are $o_p(1)$ by Lemmas 6 and 7.

Now consider Assumption 1(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

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

by Assumption 3(ii)-(iv), where $\hat{\mathbf{p}}_i = \mathbf{x}_i/C_i$. Moreover, with $\mathbf{p}_i = \mathbb{E}[\mathbf{x}_i/C_i|C_i,\boldsymbol{w}_i]$, we have

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

where 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, the third is by Lemma 4 in Appendix C and independence of $\mathbf{x}_i$ and $\mathbf{q}_i$ conditional on $(C_i,\boldsymbol{w}_i)$, and the fourth is by (12) and independence of C_i and $(Y_i,\mathbf{q}_i,\boldsymbol{w}_i)$. Therefore, $\frac{1}{\sqrt{n}}\sum_{i=1}^n(\hat{\mathbf{p}}_i-\mathbf{p}_i)\varepsilon_i \to_p \mathbf{0}$ 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 3(v) that

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

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

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

Evidently, $T_{1,n} \to_p \mathbb{E}\left[\varepsilon_i^2 \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]$ by Assumption 3(v). 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 3(ii)-(v). Since $\mathbf{p}_i, \hat{\mathbf{p}}_i \in \Delta^{V-1}$, 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 in Appendix C and Assumption 3(v). 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 3(ii)-(v). 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 in Appendix C and Assumption 3(v). Therefore, $T_{2,n} \to_p \mathbf{0}$.

Now consider $T_{3,n}$. We have $\hat{\varepsilon}_i - \varepsilon_i = \hat{\boldsymbol{\xi}}_i^T(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) + (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T\boldsymbol{\gamma}$, where

$$\begin{aligned}\max_{1\leq i\leq n}\left|\hat{\boldsymbol{\xi}}_i^T(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})\right| &\leq \left(\max_{1\leq i\leq n}\|\mathbf{q}_i\|\right)\|\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}\| \\ &\quad + \left(\max_{1\leq i\leq n}\|(\hat{\boldsymbol{\theta}}_i - \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\hat{\mathbf{p}}_i)\| + \|\boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\|\right)\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\| \to_p 0,\end{aligned}$$

where the first term is 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 3(v), and the second term follows by Assumption 3(iv), consistency of $\hat{\boldsymbol{\gamma}}$, $\|\hat{\mathbf{p}}_i\| \leq 1$, and because $\|(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}\| = O_p(1)$ by Assumption 3(ii)-(iii). Moreover,

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

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 3(ii)-(iv), and the third converges in probability to zero by Lemma 8. Hence, $\max_{1\le i\le n}|\hat{\varepsilon}_i-\varepsilon_i|\to_p 0$.

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

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

because $\frac{1}{n}\sum_{i=1}^n|\varepsilon_i|\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T=O_p(1)$ by control of $T_{1,n}$ and $T_{2,n}$, which imply $\frac{1}{n}\sum_{i=1}^n\varepsilon_i^2\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T=O_p(1)$, and $\frac{1}{n}\sum_{i=1}^n\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T=O_p(1)$ by Lemmas 6 and 7, and Assumption 3(v). ∎

Proof of Theorem 4. We first prove part 1. In the proof of Theorem 1, it was shown that $\left(\frac{1}{n}\sum_{i=1}^n\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1}\to_p\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]^{-1}$. Hence, by consistency of $\hat{\kappa}$ and $\hat{\boldsymbol{\Omega}}$, we have that

$$\hat{\kappa}\left(\frac{1}{n}\sum_{i=1}^n\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1}\begin{bmatrix}\hat{\boldsymbol{\Omega}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\to_p\kappa\,\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]^{-1}\begin{bmatrix}\boldsymbol{\Omega} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}. \tag{17}$$

Since the matrix on the right-hand side is finite, we have that

$$\frac{\hat{\kappa}}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^n\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1}\begin{bmatrix}\hat{\boldsymbol{\Omega}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}=O_p(n^{-1/2}),$$

and so

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

because $(\mathbf{I}-\mathbf{A})^{-1} = \mathbf{I} + \mathbf{A} + O(\|\mathbf{A}\|^2)$ as $\|\mathbf{A}\| \to 0$ and the inverse exists with probability approaching one. Post-multiplying both sides by $\hat{\boldsymbol{\psi}}$ gives that $\hat{\boldsymbol{\psi}}^{bcm} = \hat{\boldsymbol{\psi}}^{bca} + O_p(n^{-1})$.

Since both bias-corrected estimators are first-order asymptotically equivalent, it suffices to analyze $\hat{\boldsymbol{\psi}}^{bca}$. We have

$$\sqrt{n}(\hat{\boldsymbol{\psi}}^{bca} - \boldsymbol{\psi}) = \sqrt{n}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) + \hat{\kappa} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T \right)^{-1} \begin{bmatrix} \hat{\boldsymbol{\Omega}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \hat{\boldsymbol{\psi}}.$$

The first term is asymptotically normal with mean and variance given by (10). For the second term, Theorem 1 implies $\hat{\boldsymbol{\psi}} \to_p \boldsymbol{\psi}$, so it follows by (17) that

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

Combining with (10) and using the continuous mapping theorem, we conclude that

$$\sqrt{n}(\hat{\boldsymbol{\psi}}^{bca} - \boldsymbol{\psi}) \to_d N(\mathbf{0}, \mathbf{V}),$$

as required.

Part 2. now follows from part 1., consistency of $\hat{\mathbf{V}}$, which was established in Theorem 1, and positive-definiteness of $\mathbf{V}$. ∎

Proof of Lemma 1. Let $FP_i := \mathbb{E}[\hat{\theta}_i(1-\theta_i)|\mathbf{x}_i, \mathbf{p}_i]$. Since $\sqrt{n}\,\mathbb{E}[\hat{\theta}_i(1-\theta_i)] \to \kappa > 0$, it is enough to show that

$$\frac{\widehat{FPR}}{\frac{1}{m}\sum_{i=1}^{m} FP_i} \to_p 1, \tag{18}$$

and

$$\frac{\frac{1}{m}\sum_{i=1}^{m} FP_i}{\mathbb{E}[\hat{\theta}_i(1-\theta_i)]} \to_p 1. \tag{19}$$

We first show (19). By Chebyshev's inequality, with probability at least $1 - C^{-2}$ we have

$$\left| \sum_{i=1}^{m} FP_i - m\,\mathbb{E}[\hat{\theta}_i(1-\theta_i)] \right| \leq C \sqrt{m\,\mathbb{E}[(\hat{\theta}_i(1-\theta_i))^2]},$$

for any $C > 0$. But $\mathbb{E}[(\hat{\theta}_i(1-\theta_i))^2] < \mathbb{E}[\hat{\theta}_i(1-\theta_i)]$ because $0 \leq \hat{\theta}_i(1-\theta_i) \leq 1$. Moreover, the conditions $\sqrt{n}\,\mathbb{E}[\hat{\theta}_i(1-\theta_i)] \to \kappa > 0$ and $n/m^2 \to 0$ together imply $m\,\mathbb{E}[\hat{\theta}_i(1-\theta_i)] \to +\infty$. Setting $C = \epsilon_n \sqrt{m\,\mathbb{E}[\hat{\theta}_i(1-\theta_i)]}$ with $0 < \epsilon_n < \frac{1}{2}$ and $\epsilon_n \to 0$ sufficiently slowly that $C \to \infty$,

we deduce that

$$\left| \frac{\sum_{i=1}^m FP_i}{m\,\mathbb{E}[\hat{\theta}_i(1-\theta_i)]} - 1 \right| \leq \epsilon_n \tag{20}$$

holds with probability approaching one. This proves (19).

Now consider (18). 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

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

Letting $\mathcal{A}_n$ denote the event upon which (20) holds, we then have

$$\begin{aligned}
\Pr\left(\left| \frac{\widehat{FPR}}{\frac{1}{m}\sum_{i=1}^m FP_i} - 1 \right| > \delta \right) &\leq \mathbb{E}\left[\Pr\left(\left| \frac{\widehat{FPR}}{\frac{1}{m}\sum_{i=1}^m FP_i} - 1 \right| > \delta \middle| (\mathbf{x}_i, \mathbf{q}_i)_{i=1}^m \right) \mathbb{I}[(\mathbf{x}_i, \mathbf{q}_i)_{i=1}^m \in \mathcal{A}_n)] \right] \\
&\quad + \Pr((\mathbf{x}_i, \mathbf{q}_i)_{i=1}^m \in \mathcal{A}_n^c) \\
&\leq 2e^{-\delta^2(1-\epsilon_n) m\,\mathbb{E}[\hat{\theta}_i(1-\theta_i)]/3} + \Pr((\mathbf{x}_i, \mathbf{q}_i)_{i=1}^m \in A_n^c) \to 0,
\end{aligned}$$

since $\epsilon_n \to 0$ and $m\,\mathbb{E}[\hat{\theta}_i(1-\theta_i)] \to +\infty$. ∎

Proof of Lemma 2. Consistency of $\hat{\mathbf{\Omega}}$ 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, $\hat{\kappa} \to_p \kappa$. ∎

References

Acosta, M., Brennan, C. M., and Jacobson, M. M. (2024). Constructing high-frequency monetary policy surprises from SOFR futures. *Economics Letters*, 242:111873.

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.

Adams-Prassl, A., Waters, T., Balgova, M., and Qian, M. (2023). Firm concentration & job design: The case of schedule flexible work arrangements. Technical report, Institute for Fiscal Studies.

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. Working Paper 30446, NBER.

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

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023a). Prediction-powered inference. *Science*, 382(6671):669–674.

Angelopoulos, A. N., Duchi, J. C., and Zrnic, T. (2023b). PPI++: Efficient Prediction-Powered Inference. *arXiv:2311.01453 [stat.ML]*.

Argyle, B., Indarte, S., Iverson, B., and Palmer, C. (2025). Racial Disparities and Bias in Consumer Bankruptcy. Working Paper 33575, NBER.

Ash, E., Morelli, M., and Vannoni, M. (2025). More Laws, More Growth? Evidence from U.S. States. *Journal of Political Economy*, (ja).

Avivi, H. (2024). Are Patent Examiners Gender Neutral? Working Paper, UC Berkeley.

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. Working Paper 28731, NBER.

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.

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*, 3:993–1022.

Blinder, A. (2018). Through a crystal ball darkly: The future of monetary policy communication. *AEA Papers and Proceedings*, 108:567–571.

Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., and Jansen, D.-J. (2008). Central

bank communication and monetary policy: A survey of theory and evidence. *Journal of Economic Literature*, 46(4):910–45.

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

Bursztyn, L., Chaney, T., Hassan, T. A., and Rao, A. (2024). The Immigrant Next Door. *American Economic Review*, 114(2):348–384.

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

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

Caldara, D. and Iacoviello, M. (2022). Measuring Geopolitical Risk. *American Economic Review*, 112(4):1194–1225.

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. (2025). Demand Estimation with Text and Image Data. *arXiv:2503.20711 [econ.GN]*.

Council of Economic Advisers (2025). Economic Report of the President, 2025.

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

Draca, M. and Schwarz, C. (2021). How Polarized are Citizens? Measuring Ideology from the Ground-Up. Technical Report ID 3154431, SSRN.

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.

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

Einav, L., Finkelstein, A., and Mahoney, N. (2024). Producing Health: Measuring Value Added of Nursing Homes. Working Paper 30228, NBER.

Evdokimov, K. S. and Zeleneev, A. (2023). Simple Estimation of Semiparametric Models with Measurement Errors. *arXiv:2306.14311 [econ.EM]*.

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. Technical Report WP 25-14, Federal Reserve Bank of Philadelphia.

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.

Goldsmith-Pinkham, P. and Shue, K. (2023). The Gender Gap in Housing Returns. *The Journal of Finance*, 78(2):1097–1145.

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.

Gürkaynak, R. S., Sack, B., and Swanson, E. (2005). Do actions speak louder than words? The response of asset prices to monetary policy actions and statements. *International Journal of Central Banking*, 1:55–93.

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. Working Paper 31007, NBER.

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.

Ke, S., Olea, J. L. M., and Nesbit, J. (2024). Robust Machine Learning Algorithms for Text Analysis. *Quantitative Economics*, 15(4):939–970.

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

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

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.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2025). Large Language Models: An Applied Econometric Framework. *arXiv:2412.07031 [econ]*.

Magnolfi, L., McClure, J., and Sorensen, A. (2025). Triplet Embeddings for Demand Estimation. *American Economic Journal: Microeconomics*, 17(1):282–307.

Mardia, J., Jiao, J., Tánczos, E., Nowak, R. D., and Weissman, T. (2019). Concentration Inequalities for the Empirical Distribution. *arXiv:1809.06522 [cs.IT]*.

Miao, J. and Lu, Q. (2024). Task-Agnostic Machine-Learning-Assisted Inference. *arXiv:2405.20039 [stat.ML]*.

Modarressi, I., Spiess, J., and Venugopal, A. (2025). Causal Inference on Outcomes Learned from Text. *arXiv:2503.00725 [econ.EM]*.

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.

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.

Rambachan, A., Singh, R., and Viviano, D. (2025). Program Evaluation with Remotely Sensed Outcomes. *arXiv:2411.10959 [econ.EM]*.

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.

Sacher, S., Battaglia, L., and Hansen, S. (2024). Hamiltonian Monte Carlo for Regression with High-Dimensional Categorical Data. *arXiv:2107.08112 [econ.EM]*.

Sanford, L. C., Ayers, M., Gordon, M., and Stone, E. (2025). Adversarial Debiasing for Unbiased Parameter Recovery. *arXiv:2502.12323 [cs.LG]*.

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]*.

Schennach, S. (2022). Measurement Systems. *Journal of Economic Literature*, 60(4):1223–1263.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2022). Measuring news sentiment. *Journal of Econometrics*, 228(2):221–243.

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, J. C. and Xia, F. D. (2016). Measuring the macroeconomic impact of monetary policy at the zero lower bound. *Journal of Money, Credit and Banking*, 48(2-3):253–291.

Wu, L. and Yang, T.-T. (2024). Behavioral Responses to Estate Taxation: Evidence from Taiwan. Technical report, University College London.

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*.

Zrnic, T. and Candès, E. J. (2024). Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121.

Supplemental 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

April 29, 2025

B Additional Results and Discussion

B.1 Fixed-DGP Asymptotics

For completeness, here we study 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 fixed-DGP framework approximates empirical settings with a large number of observations and a large amount of measurement error per observation.

Assumption 4. *(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) $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T \to_p \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]$, $\frac{1}{n}\sum_{i=1}^n (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T \to_p \mathbf{H}$ *with* $\mathbf{H}$ *a finite non-random symmetric matrix,* $\frac{1}{n}\sum_{i=1}^n (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\boldsymbol{\theta}_i^T \to_p \mathbf{G}$ *with* $\mathbf{G}$ *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 4(i) is standard. Assumption 4(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 4(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. Let $\boldsymbol{\Delta} = \mathbf{H} + \mathbf{G} + \mathbf{G}^T$.

Theorem 5. *Suppose that Assumption 4 holds. Then*

$$\hat{\boldsymbol{\psi}} \to_p \left(\mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right] + \begin{bmatrix} \boldsymbol{\Delta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right)^{-1} \left(\mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right] + \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right) \boldsymbol{\psi}, \tag{21}$$

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

$$\operatorname{plim}(\hat{\boldsymbol{\psi}}) = \boldsymbol{\psi} - \mathbb{E}\left[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right]^{-1} \begin{bmatrix} (\mathbf{H} + \mathbf{G}^T)\boldsymbol{\gamma} \\ \mathbf{0} \end{bmatrix} + O\left(\|\boldsymbol{\Delta}\| \max\{\|\boldsymbol{\Delta}\|, \|\mathbf{G}\|\}\right). \tag{22}$$

Theorem 5 shows that $\hat{\boldsymbol{\psi}}$ is inconsistent due to measurement error in $\hat{\boldsymbol{\theta}}_i$. More constructively, Theorem 5 shows that the measurement error bias is, to first order, proportional to the precision of $\boldsymbol{\theta}_i$. The matrix $\mathbf{H}$ represents the variance of measurement error, while $\mathbf{G}$ represents the covariance of measurement error and $\boldsymbol{\theta}_i$. In many cases, measurement error is "classical" ($\mathbf{H} > \mathbf{0}$, $\mathbf{G} = \mathbf{0}$), but in "non-classical" settings, such as latent binary labels (Aigner 1973), $\mathbf{G} \neq \mathbf{0}$. Expressions for $\mathbf{H}$ and $\mathbf{G}$ in the context of AI/ML-generated labels and topic models are derived in the following subsections.

B.2 AI/ML-Generated Labels

Here we first generalize the basic framework from Section 4.1.1 to allow for multiple categories and randomized classifiers. Let the vector $\boldsymbol{\theta}_i = (\theta_{i,k})_{k=1}^K$ indicate membership of $K+1$ distinct categories labeled $0, 1, \ldots, K$. 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 $\mathbf{p}(\mathbf{x}_i) = (p_k(\mathbf{x}_i))_{k=1}^K$. If $\mathbf{q}_i$ is relevant for predicting $\boldsymbol{\theta}_i$, then we implicitly treat $\mathbf{q}_i$ as a component of $\mathbf{x}_i$ to simplify notation.

For the classifier, 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 $\mathbf{r}(\mathbf{x}_i, U_i)|\mathbf{x}_i \sim \text{Multinomial}(1, \boldsymbol{\pi}(\mathbf{x}_i))$. Here $\boldsymbol{\pi}(\mathbf{x}_i) = (\pi_k(\mathbf{x}_i))_{k=1}^K$, where $\pi_k(\mathbf{x}_i)$ denotes the probability that the classifier assigns label k given $\mathbf{x}_i$. This nests deterministic classifiers, where $\mathbf{r}(\mathbf{x}_i, U_i) = \mathbf{r}(\mathbf{x}_i) = \boldsymbol{\pi}(\mathbf{x}_i)$, with $\pi_k(\mathbf{x}_i) = 1$ for at most one $k \in \{1, \ldots, K\}$ (with k depending on $\mathbf{x}_i$) and $\pi_j(\mathbf{x}_i) = 0$ for all $j \neq k$.

B.2.1 Fixed-DGP Asymptotics

We first provide primitive conditions for the fixed-DGP case and derive expressions for the corresponding asymptotic bias. 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 5. *(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) $\mathbb{E}\left[\left(\boldsymbol{\pi}(\mathbf{x}_i) - \mathbf{p}(\mathbf{x}_i)\right) \mathbf{q}_i^T\right] = \mathbf{0}$.
(iv) $\mathbb{E}\left[\boldsymbol{\pi}(\mathbf{x}_i)\varepsilon_i\right] = \mathbf{0}$.

Assumption 5(i) imposes minimal structure on the AI/ML-generated predictions $\hat{\boldsymbol{\theta}}_i$. It allows the classifier to be pre-trained, in which case one should interpret the analysis as holding conditional on the training sample. Assumption 5(ii) imposes standard moment and rank conditions. Assumption 5(iii) requires the classifier's prediction errors to be uncorrelated with the controls $\mathbf{q}_i$. As discussed in the main text, it is straightforward to relax this condition without changing our main point. Finally, Assumption 5(iv) says that the regression errors ε_i are uncorrelated with $\boldsymbol{\pi}(\mathbf{x}_i)$. A sufficient conditions is $\mathbb{E}[\varepsilon_i|\mathbf{x}_i, \mathbf{q}_i] = 0$.

Theorem 6. *Suppose that Assumption 5 holds. Then Assumption 4 holds and the OLS estimator $\hat{\boldsymbol{\psi}}$ has probability limit given by (21), with*

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

If $\mathbf{H}$ *and* $\mathbf{G}$ *are small, then first-order bias is proportional to*

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

B.2.2 Sequence-of-DGPs Asymptotics

Consider the matrix $\mathbf{H}$ from Theorem 6. Misclassification rates for each of the K labels are collected along the diagonal of $\mathbf{H}$:

$$(\mathbf{H})_{k,k} = \mathbb{E}\left[\pi_k(\mathbf{x}_i) + p_k(\mathbf{x}_i) - 2\pi_k(\mathbf{x}_i)p_k(\mathbf{x}_i)\right], \quad k = 1, \ldots, K.$$

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

$$\operatorname{tr}(\mathbf{H}) \to 0 \tag{23}$$

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 $FP(\mathbf{x}_i) = \sum_{k=1}^{K} FP_k(\mathbf{x}_i)$ 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

$$\lim_{n\to\infty} \sqrt{n}\, \mathbb{E}\left[\operatorname{diag}(\boldsymbol{\pi}(\mathbf{x}_i)) - \boldsymbol{\pi}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^T\right] = \kappa\, \boldsymbol{\Omega}, \tag{24}$$

where

$$\lim_{n\to\infty} \sqrt{n}\,\mathbb{E}\left[FP(\mathbf{x}_i)\right] = \kappa \geq 0, \tag{25}$$

and

$$\lim_{n\to\infty} \frac{\mathbb{E}\left[\text{diag}(\boldsymbol{\pi}(\mathbf{x}_i)) - \boldsymbol{\pi}(\mathbf{x}_i)\mathbf{p}(\mathbf{x}_i)^T\right]}{\mathbb{E}\left[FP(\mathbf{x}_i)\right]} = \boldsymbol{\Omega}, \tag{26}$$

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, $\kappa > 0$ allows the total false-positive rate to be 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.

Assumption 6. *(i) Conditions (23)-(26) hold.*
(ii) $\sqrt{n}\max_{1\leq i\leq n}\|\hat{\boldsymbol{\theta}}_i - \mathbf{r}(\mathbf{x}_i, U_i)\| \to_p 0$.
(iii) $\mathbb{E}\left[\|\mathbf{q}_i\|^4\right] < \infty$, $\mathbb{E}\left[\varepsilon_i^4\right] < \infty$, *and* $\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]$ *has full rank.*
(iv) $\sqrt{n}\,\mathbb{E}\left[\left(\boldsymbol{\pi}(\mathbf{x}_i) - \mathbf{p}(\mathbf{x}_i)\right)\mathbf{q}_i^T\right] \to \mathbf{0}$.
(v) $\sqrt{n}\,\mathbb{E}\left[\boldsymbol{\pi}(\mathbf{x}_i)\varepsilon_i\right] \to \mathbf{0}$.

Assumption 6(i) formalizes the asymptotic framework. Assumption 6(ii) slightly strengthens Assumption 5(i) to require convergence at a faster-than-root-n rate. Assumption 6(iii) is standard. Assumption 6(iv) and 6(v) weaken Assumptions 5(iii) and 5(iv). As before, it is possible to relax these conditions without changing our main point.

Theorem 7. *Suppose that Assumption 6 holds. Then Assumption 1 holds and the OLS estimator* $\hat{\boldsymbol{\psi}}$ *has asymptotic distribution given by (10) with* κ *given in (25) and* $\boldsymbol{\Omega}$ *given in (26). Moreover, two-step standard errors are consistent.*

B.3 Topic Models

Section 4.1.2 presented a set of results for the sequence-of-DGPs asymptotic framework. Here we present a complementary set of results for the fixed-DGP case, before turning to a discussion of identification of the topic model in our setting.

B.3.1 Fixed-DGP Asymptotics

As in the main text, we implicitly 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 sampling error and regression errors are uncorrelated. These assumptions seem very reasonable and can be relaxed: doing so simply complicates the expressions

below. We also implicitly require that $\mathbb{E}[\varepsilon_i(\boldsymbol{w}_i, \mathbf{q}_i)] = \mathbf{0}$. That is, no relevant topic weights have been omitted from the regression. 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 n becomes large.

Assumption 7. *(i)* $\mathbf{B}$ *has full rank.*
(ii) $\hat{\mathbf{B}} \to_p \mathbf{B}$.
(iii) $\max_{1 \le i \le n} \|\hat{\boldsymbol{\theta}}_i - \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}(\mathbf{x}_i/C_i)\| \to_p 0$.
(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 7(i) says that none of the topics are redundant. Assumption 7(ii) says that $\hat{\mathbf{B}}$ is consistent, which is satisfied by many estimators for topic models including those of Bing et al. (2020), Wu et al. (2023), and Ke and Wang (2022). Assumption 7(iii) imposes some structure on the $\hat{\boldsymbol{\theta}}_i$. This condition is not vacuous: $\boldsymbol{\theta}_i = \boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\mathbb{E}[\mathbf{x}_i/C_i|C_i, \boldsymbol{w}_i]$ by Assumption 7(i), so one could set $\hat{\boldsymbol{\theta}}_i = \boldsymbol{S}(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}}(\mathbf{x}_i/C_i)$. Assumption 7(iv) is standard.

Theorem 8. *Suppose that Assumption 7 holds. Then Assumption 4 holds and the OLS estimator* $\hat{\boldsymbol{\psi}}$ *has probability limit given by (21) with*

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

B.3.2 Identification

Recall that we are sampling $(\mathbf{x}_i, C_i)$ according to the topic model (6) in either a fixed-DGP setting (Appendix B.3.1) or from a sequence of DGPs in which the distribution of $(\mathbf{x}_i, \boldsymbol{w}_i)|C_i$ is fixed but the distribution of C_i is changing with n so that (12) holds (Section 4.1.2). In either case, the sampling framework is one in which the number of observations n, and hence the number of $\boldsymbol{w}_i$, is increasing.

This setting differs from recent works in econometrics that have studied identification. Ke et al. (2024) and Freyaldenhoven et al. (2023) consider a fixed-n setting where each $C_i \to \infty$ so that $\frac{\mathbf{x}_i}{C_i} \to_p \mathbf{p}_i := \mathbb{E}[\frac{\mathbf{x}_i}{C_i}|C_i, \boldsymbol{w}_i]$, the vector of multinomial probabilities for observation i. In a fixed-n setting, identification concerns uniqueness of the factorization $\mathbf{P} = \mathbf{B}^T\boldsymbol{W}$ with $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_n] \in (\Delta^{V-1})^n$, $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n] \in (\Delta^{K-1})^n$ and $\mathbf{B} \in (\Delta^{V-1})^K$, conditional on the n (fixed) sampled units.

Within a fixed-n context, identification of $\mathbf{B}$ is commonly achieved in text applications by assuming the existence of anchor words that are known to appear in some topics but not others. In essence, these amount to zero restrictions on elements of $\mathbf{B}$.

In the fixed-DGP case we consider in Appendix B.3.1, anchor words are also sufficient for $\mathbf{B}$ to be consistently estimable as $n \to \infty$ and thus, by construction, identified. See,

e.g., Bing et al. (2020), Wu et al. (2023), and Ke and Wang (2022). The same is true for the sequence-of-DGPs case we consider in Section 4.1.2. But, in that case, we are in effect sampling each true multinomial probability $\mathbf{p}_i$, since each C_i is growing with n. Let $\mathcal{P} \subseteq \Delta^{V-1}$ denote the support of $\mathbf{p}_i$.[26] Then $\mathbf{B}$ is identified if it is the unique element of $(\Delta^{V-1})^K$ for which $\{(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\mathbf{p} : \mathbf{p} \in \mathcal{P}\} \subseteq \Delta^{K-1}$. Anchor words are sufficient for this, but not necessary when the $\boldsymbol{w}_i$ have rich support.

To see this, consider the following illustration within the context of the AI/ML-generated index running example (Application 3 in Section 3). Recall that here we have

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

Suppose θ_i has probability density function that is strictly positive on $(0,1)$. Then $\mathcal{P}$ is the convex hull of $[\beta_1, 1-\beta_1]^T$ and $[\beta_0, 1-\beta_0]^T$. These extreme values of $\mathcal{P}$ identify β_1 and β_0, and therefore $\mathbf{B}$ is identified without anchor words.

C 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.

C.1 Fixed-DGP Asymptotics

Proof of Theorem 5. First consider the denominator. We have

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

[26] In the fixed-n case which conditions on the n observed units, we have $\mathcal{P} = \{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$.

Hence,

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

by Assumption 4(ii). The right-hand side is finite by Assumption 4(i) and invertible by assumption.

For the numerator term, we have

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

Hence,

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

by Assumption 4(ii). The first result 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. ∎

C.2 AI/ML-Generated Labels

Proof of Theorem 6. Assumption 4(i) holds by Assumption 5(ii) and the fact that $\|\boldsymbol{\theta}_i\| \leq 1$. The first part of Assumption 4(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 5(i) and the law of large numbers,

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

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 5(i)-(ii). Then by Assumption 5(ii)-(iii) and the law of large numbers,

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

For the final part of Assumption 4(ii), first note

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \varepsilon_i = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\theta}}_i \varepsilon_i \\ \frac{1}{n} \sum_{i=1}^{n} \mathbf{q}_i \varepsilon_i \end{bmatrix}, \tag{27}$$

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(ii). Moreover,

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

by Assumption 5(i)-(ii). Finally,

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

by the law of large numbers, independence of $\mathbf{r}(\mathbf{x}_i, U_i)$ and ε_i conditional on $\mathbf{x}_i$ (for the first equality), and Assumption 5(iv) (for the second equality). ∎

Lemma 3. *Let $\mathbf{z}_i$ be a random vector with finite fourth moment and let Assumption 6(ii) and (23) 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 6(ii) 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 $\mathbb{E}[\|\mathbf{z}_i\|^4] < \infty$ and $r_k(\mathbf{x}_i, U_i) - \theta_{i,k}$ takes values in $\{-1, 0, 1\}$, and convergence to zero is by (23) because $\mathbb{E}\left[\left(r_k(\mathbf{x}_i, U_i) - \theta_{i,k}\right)^2\right] = (\mathbf{H})_{k,k}$. The result now follows by Chebyshev's inequality. ∎

Proof of Theorem 7. Assumption 1(i) holds by Assumption 6(iii) and the fact that $\|\boldsymbol{\theta}_i\| \leq 1$.

We now verify Assumption 1(ii). First by Lemma 3 and Assumption 6, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\mathbf{q}_i^T = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left((\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\mathbf{q}_i^T - \mathbb{E}[(\boldsymbol{\pi}(\mathbf{x}_i) - \mathbf{p}(\mathbf{x}_i))\mathbf{q}_i^T]\right) \to_p \mathbf{0}, \tag{28}$$

which establishes the final part of Assumption 1(ii). Similarly, by Assumption 6(ii),

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

Hence, it follows by Lemma 3 and (24) that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T \to_p \kappa\,\boldsymbol{\Omega}, \tag{29}$$

which establishes the second part of Assumption 1(ii). For the first part of Assumption 1(ii), we note

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

Displays (28) and (29) together imply 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 6(iii). It therefore remains to show that the second term on the right-hand side of the above display is asymptotically negligible. In view of (28) it is enough to show

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

By Lemma 3,

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

where $\mathbb{E}\left[\boldsymbol{\theta}_i(\mathbf{r}(\mathbf{x}_i,U_i)-\boldsymbol{\theta}_i)^T\right]=\mathbb{E}\left[\mathbf{p}(\mathbf{x}_i)\boldsymbol{\pi}(\mathbf{x}_i)^T-\mathrm{diag}(\mathbf{p}(\mathbf{x}_i))\right]\to\mathbf{0}$ by (23), noting the diagonal entries are $\mathbb{E}[p_k(\mathbf{x}_i)(\pi_k(\mathbf{x}_i)-1)]$ where $0\le\mathbb{E}[p_k(\mathbf{x}_i)(1-\pi_k(\mathbf{x}_i))]\le(\mathbf{H}_{k,k})\to 0$ and the off-diagonals are $\mathbb{E}[p_k(\mathbf{x}_i)\pi_j(\mathbf{x}_i)]$, where $\sum_{j\neq k}\mathbb{E}[p_k(\mathbf{x}_i)\pi_j(\mathbf{x}_i)]\le\mathbb{E}[p_k(\mathbf{x}_i)(1-\pi_k(\mathbf{x}_i))]\to 0$. This completes the verification of Assumption 1(ii).

Now consider Assumption 1(iii). For the first part, we have

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

Note that $\sqrt{n}\,\mathbb{E}\left[(\mathbf{r}(\mathbf{x}_i,U_i)-\boldsymbol{\theta}_i)\varepsilon_i\right]=\sqrt{n}\,\mathbb{E}\left[\boldsymbol{\pi}(\mathbf{x}_i)\varepsilon_i\right]\to\mathbf{0}$ by Assumption 6(v) and the fact that $\mathbb{E}[\varepsilon_i\boldsymbol{\theta}_i]=\mathbf{0}$. Hence, by Lemma 3 using Assumption 6(ii)-(iii), 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 1(iii) now holds by the central limit theorem and Assumption 6(iii).

To complete the verification of Assumption 1(iii), 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 $T_{1,n}\to_p\mathbb{E}\left[\varepsilon_i^2\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]$ by Assumption 6(iii). To show $T_{2,n}\to_p\mathbf{0}$, it suffices to show

$$T_{2,a,n}:=\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\left(\hat{\boldsymbol{\theta}}_i\hat{\boldsymbol{\theta}}_i^T-\boldsymbol{\theta}_i\boldsymbol{\theta}_i^T\right)\to_p\mathbf{0},\tag{30}$$

and

$$T_{2,b,n} := \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \mathbf{q}_i \left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right)^T \to_p \mathbf{0}. \tag{31}$$

For $T_{2,a,n}$, we may first deduce by Assumption 6(ii)-(iii) 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 $\boldsymbol{\theta}_i \boldsymbol{\theta}_i^T = \mathrm{diag}(\boldsymbol{\theta}_i)$ 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 6(iii) and the second term is $o_p(1)$ because $\mathbb{E}\left[\|\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i\|^2\right] = \mathrm{tr}(\mathbf{H}) \to 0$ by (23), proving (30). For $T_{2,b,n}$, we may similarly deduce by Assumption 6(ii)-(iii) 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 6(iii). For the third term, note that $\|\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i\|^4 \leq 4\|\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i\|^2$ because $\|\mathbf{r}(\mathbf{x}_i, U_i) - \boldsymbol{\theta}_i\|^2 \in \{0, 1, 2\}$. Hence, the third term is $o_p(1)$ by (23), proving (31).

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

$$T_{3,a,n} := \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{\xi}_i^T \boldsymbol{\psi} - \hat{\boldsymbol{\xi}}_i^T \hat{\boldsymbol{\psi}})^2 \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T \to_p \mathbf{0}, \tag{32}$$

and

$$T_{3,b,n} := \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i (\boldsymbol{\xi}_i^T \boldsymbol{\psi} - \hat{\boldsymbol{\xi}}_i^T \hat{\boldsymbol{\psi}}) \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T \to_p \mathbf{0}. \tag{33}$$

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

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

where $\frac{1}{n}\sum_{i=1}^{n}\|\hat{\boldsymbol{\xi}}_i\|^4 = O_p(1)$ and $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^4 = O_p(1)$ by Assumption 6(ii)-(iii). Moreover, $\boldsymbol{\xi}_i^T\boldsymbol{\psi} - \hat{\boldsymbol{\xi}}_i^T\hat{\boldsymbol{\psi}} = (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T\boldsymbol{\gamma} + \hat{\boldsymbol{\xi}}_i^T(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}})$. 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 6(ii). 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 (32) and (33). ∎

C.3 Topic Models

Throughout this section and the next, let $\hat{\mathbf{p}}_i = \frac{\mathbf{x}_i}{C_i}$ and $\mathbf{p}_i = \mathbb{E}[\frac{\mathbf{x}_i}{C_i}|C_i, \boldsymbol{w}_i]$. The next two lemmas apply in both fixed-DGP and sequence-of-DGPs frameworks.

Lemma 4. *Suppose that (6) 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 (6) 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 7(i)-(iii) hold. Then*

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

Proof of Lemma 5. In view of Assumption 7(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 7(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 7(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 $\boldsymbol{w}_i$. ∎

Proof of Theorem 8. Assumption 4(i) holds by Assumption 7(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 4(ii) holds because $\mathbb{E}[\|\boldsymbol{\xi}_i\|^2] < \infty$. For the second part, we have

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

by Lemma 5. Further, $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\theta}_i \boldsymbol{\theta}_i^T \to_p \mathbb{E}\left[\boldsymbol{\theta}_i \boldsymbol{\theta}_i^T\right]$. To complete the proof of the second, third, and fourth parts of Assumption 4(ii), it suffices to show that

$$\frac{1}{n}\sum_{i=1}^{n} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\boldsymbol{\xi}_i^T \to_p \mathbf{0}. \tag{34}$$

To this end, in view of Assumption 7(iii)-(iv), we have

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

Note $(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}} \to_p (\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}$ by Assumption 7(i)-(ii), and $\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{p}}_i \boldsymbol{\xi}_i^T = O_p(1)$ by Assumption 7(iv) and the fact that $\|\hat{\mathbf{p}}_i \boldsymbol{\xi}_i^T\| \leq \|\boldsymbol{\xi}_i\|$. Hence,

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

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 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 (34), from which we also conclude that $\mathbf{G} = \mathbf{0}$.

To verify the final part of Assumption 4(ii), first note

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \varepsilon_i = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\theta}}_i \varepsilon_i \\ \frac{1}{n}\sum_{i=1}^{n} \mathbf{q}_i \varepsilon_i \end{bmatrix}, \tag{35}$$

where $\frac{1}{n}\sum_{i=1}^{n} \mathbf{q}_i \varepsilon_i \to_p \mathbf{0}$ by Assumption 7(iv). Moreover, by Assumption 7(iii)-(iv), we have

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

We also have $(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}\hat{\mathbf{B}} \to_p (\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}$ by Assumption 7(i)-(ii). Moreover, Assumption 7(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 independence of $(\mathbf{x}_i, C_i)$ and ε_i conditional on $(\boldsymbol{w}_i, \mathbf{q}_i)$, and $\mathbb{E}[\varepsilon_i(\boldsymbol{w}_i, \mathbf{q}_i)] = \mathbf{0}$. ∎

C.4 Additional Results for the Proof of Theorem 3

Lemma 6. *Let Assumption 3(i)-(iv) 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 3(iv), 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 3(ii)-(iii) 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

$$\begin{aligned}\mathbb{E}\left[\hat{\mathbf{p}}_i\left(\hat{\mathbf{p}}_i - \mathbf{p}_i\right)^T\right] &= \mathbb{E}\left[\left(\hat{\mathbf{p}}_i - \mathbf{p}_i\right)\left(\hat{\mathbf{p}}_i - \mathbf{p}_i\right)^T\right] \\ &= \mathbb{E}\left[\frac{1}{C_i}\right]\left(\mathrm{diag}\left(\mathbf{B}^T\mathbb{E}\left[\boldsymbol{w}_i\right]\right) - \mathbf{B}^T\mathbb{E}\left[\boldsymbol{w}_i\boldsymbol{w}_i^T\right]\mathbf{B}\right). \qquad (36)\end{aligned}$$

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,

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

where the second inequality is because $\hat{\mathbf{p}}_i$ is in the simplex and the convergence to zero holds

in view of (12) and (36). 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 $T_{2,n} \to_p -\kappa \left(\boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\, \mathrm{diag}(\mathbf{B}^T \mathbb{E}[\boldsymbol{w}_i])\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\boldsymbol{S}^T - \mathbb{E}[\boldsymbol{\theta}_i \boldsymbol{\theta}_i^T] \right)$ by (36) and Assumption 3(i)-(iii) ∎

Lemma 7. *Let Assumption 3(i)-(v) 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 3(iv)-(v), where

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

Assumption 3(ii)-(iii) implies $\sqrt{n}(\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1} - \hat{\mathbf{B}}^T(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)^{-1}) \to_p \mathbf{0}$. We also have that $\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\xi}_i \mathbf{p}_i^T \| \leq \frac{1}{n} \sum_{i=1}^{n} \| \boldsymbol{\xi}_i \| = O_p(1)$, by Assumption 3(v). Hence, $T_{1,n} \to_p \mathbf{0}$. 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

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

where the first inequality is by Cauchy-Schwarz, the second is by Assumption 3(v) and the fact that $|\hat{\mathbf{p}}_{i,k} - \mathbf{p}_{i,k}| \leq 1$, and convergence to zero is by (36) and Assumption 3(i). It follows that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{X}_i \to_p \mathbf{0}$. We conclude by Assumption 3(ii)-(iii) that $T_{2,n} \to_p \mathbf{0}$. ∎

Lemma 8. *Let Assumption 3(vi) 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 $C_i\hat{\mathbf{p}}_i|(C_i, \boldsymbol{w}_i) \sim \text{Multinomial}(C_i, \mathbf{p}_i)$, for all $t > 0$ we have

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

by the union bound and Lemma 1 of Mardia et al. (2019). Then by Assumption 3(vi),

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

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. ∎