

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Economics

2024

Essays in Econometrics: Inference in Spatial Causal and Limited-Information Settings and Marginal Forecast Adjustment for Bayesian Vector Autoregressions

Cocci, Matthew David

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp01d504rp693>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

ESSAYS IN ECONOMETRICS: INFERENCE IN SPATIAL CAUSAL
AND LIMITED-INFORMATION SETTINGS AND MARGINAL
FORECAST ADJUSTMENT FOR BAYESIAN VECTOR
AUTOREGRESSIONS

MATTHEW DAVID COCCI

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
ECONOMICS

ADVISERS: ULRICH K. MÜLLER AND MIKKEL PLAGBORG-MØLLER

MAY 2024

© Copyright by Matthew David Cocci, 2024.

All Rights Reserved

Abstract

I develop and characterize valid standard errors and inference for both spatial causal inference problems and limited-information calibration, as well as define a post-estimation procedure that adjusts Bayesian vector autoregression (BVAR) forecasts given additional information in a manner consistent with full re-estimation.

Chapter 1 provides and justifies standard errors for causal inference settings with spatial assignment. I show how knowledge of and weak dependence in the assignment mechanism alone justify particular design-based standard errors and inference, avoiding restrictive assumptions on outcomes or the distribution of individual causal effects. I contrast these standard errors with common alternatives from the literature, demonstrating that these alternatives tend to exhibit overcoverage and power loss when individual treatment effects are positively spatially correlated. Finally, I extend the results to derive valid design-based standard errors and inference for a local average treatment effect (LATE) estimand under correlated treatment assignment.

Chapter 2, coauthored with Mikkel Plagborg-Møller, demonstrates how to conduct inference for calibrated model parameters when the correlation structure of moments used in calibration are unknown. Existing standard error formulas require a consistent estimate of the correlation structure. However, using only the variances of empirical moments, we derive conservative standard errors, confidence intervals, and overidentification tests that are valid even under the worst-case correlation structure. We apply the methods empirically to a model of menu cost pricing for multi-product firms and to a heterogeneous agent New Keynesian model.

Chapter 3 defines a procedure for adjusting the forecast distribution of a BVAR by exploiting additional information ex-post after estimation and states conditions under which the adjustment procedure is mean-equivalent to re-estimation of an expanded model. I show how one can apply the Frisch-Waugh-Lovell Theorem in a Bayesian context to extract the marginal information in additional predictors and use it to drive data-driven adjustments of the forecast distribution, under proper penalization as would occur under re-estimation. In an out-of-sample forecasting exercise, I compare performance of this simple adjustment procedure to a popular alternative of constructing and estimating individual, custom nowcast models for each outcome in the core BVAR. I demonstrate that the recommended procedure performs comparably for most variables.

Acknowledgements

I am deeply grateful to my advisers Ulrich Müller and Mikkel Plagborg-Møller. Over many years, they have provided unwavering support, patience, insight, and mentorship. They have always served and continue to serve as academic examples of the highest order through their teaching, work, and actions. I would also like to thank the broader econometrics faculty, especially Michal Kolesár, Bo Honoré, and Mark Watson, for their generous and thoughtful feedback.

I am also deeply grateful for the friendship and fellowship shown to me by Elisa Jácome, Mathilde Le Moigne, Vivek Sampathkumar, and Alexander Wise throughout the process. Similarly, I thank Laura Hedden for always being such a kind and supportive presence in the department.

Outside of Princeton, I am grateful to Domenico Giannone, Marco Del Negro, Marc Giannoni, and the many other unintentionally yet delightfully Italian economists with whom I have had the pleasure to work. I have benefitted tremendously from their example and advice when it comes to putting econometric theory into practice, as well as the generous investments that they have always made in aspiring and junior economists.

Finally, I am grateful to my family and many outstanding educators I have had prior to Princeton. First and foremost, I am thankful to my parents David and Theresa Cocci, as well as my partner, Kaitlin Lassen, to whom this work is dedicated. They have always been such deep and steady sources of support, patience, comfort, and motivation. Thank you also to my extended family, especially the many aunts and uncles who formed a special and supportive community for as long as I can remember (and who let me always bring a book to family events): Mario Cocci, Bob and Justine Cocci, Jeff and Cyndi Cocci, Linda and Randy Feldman, Sherry and Dave Womer, Kent and Lorraine Close, Liz and Russ Wise, and Jim and Marily Quinn. Lastly, thank you to the many wonderful, dedicated educators from St. Catharine of Siena School and Central Catholic High School, including members of the Sisters of the Immaculate Heart of Mary and the Sisters of Christian Charity. So many there and in my family have fostered a deep, foundational love of education that underpins all work.

To Mom, Dad, and Kait.

Contents

Abstract	iii
Acknowledgements	iv
1 Standard Errors under Correlated Assignment in Spatial Settings	1
1.1 Introduction	1
1.2 Framework and Preliminaries	8
1.2.1 Spatial Assignment Mechanism	9
1.2.2 Individual Units and Potential Outcomes	10
1.2.3 Estimand	12
1.2.4 Asymptotic Embedding and Inference Perspective	12
1.2.5 Analytical Example	14
1.3 Estimation of the ATE	16
1.3.1 Effect Estimator	16
1.3.2 Consistency	18
1.3.3 Asymptotic Normality	20
1.4 Variance Estimation	23
1.4.1 Estimators	24
1.4.2 Comparison of Variance Estimators	27
1.5 Results and Simulations for the Analytical Example	31
1.5.1 Weak Dependence	31
1.5.2 Comparison of Variance Estimators	33
1.5.3 Simulation Study	35
1.6 Instrumental Variables under Joint Assignment with Treatment Effect Heterogeneity	43
1.6.1 Framework	43

1.6.2	LATE Estimand	45
1.6.3	Asymptotics and Inference for LATE Estimand	46
1.7	Conclusion	49
2	Standard Errors for Calibrated Parameters	50
2.1	Introduction	50
2.2	Setup	53
2.3	Standard errors and moment selection	58
2.3.1	Worst-case standard errors and confidence intervals	58
2.3.2	Efficient moment selection	60
2.3.3	General knowledge of the covariance matrix	64
2.4	Testing	65
2.4.1	Joint testing of parameter restrictions	65
2.4.2	Over-identification testing	67
2.5	Empirical applications	68
2.5.1	Menu cost price setting in multi-product firms	68
2.5.2	Heterogeneous agent New Keynesian model	71
2.6	Conclusion	75
3	Tilting VAR Forecasts on Marginal Information	77
3.1	Introduction	77
3.2	Setup	81
3.2.1	Core and Full-Information VARX Models	81
3.2.2	Prior Specifications	84
3.2.3	Entropic Tilting	86
3.3	Tilting Procedures for Forecast Adjustment	88
3.3.1	Tilting and Re-Estimation Equivalence	88
3.3.2	Variations on the Tilting Procedure	96
3.3.3	Longer Horizon Forecasts	102
3.4	Empirical Application	103
3.4.1	Data	104
3.4.2	Modeling Specifications	107

3.4.3	Accuracy Comparison	115
3.5	Conclusion	122
A	Appendix for Chapter 1	136
A.1	Spatial Point Processes and Analytical Example	136
A.1.1	Poisson Processes	136
A.1.2	Proofs for Section 1.2.5	137
A.2	Background for Asymptotic Results	140
A.2.1	Setup and Metrics	140
A.2.2	Mixing Definitions and Relationships	142
A.2.3	Mixing Assumptions for Normality	143
A.2.4	Spatial Central Limit Theorem under Weak Dependence	145
A.2.5	Useful Covariance Inequalities using Mixing Coefficients	146
A.3	Proofs for Section 1.3	149
A.3.1	Useful Lemmas for Normality and Variance Consistency Proofs	149
A.3.2	Proofs for Section 1.3.1	150
A.3.3	Proofs for Section 1.3.2	151
A.3.4	Proofs for Section 1.3.3	153
A.3.5	Variance Estimation	158
A.4	IV Setting	174
B	Appendix for Chapter 2	177
B.1	Appendix	177
B.1.1	Technical lemmas and proofs	177
B.1.2	Simulated moments	179
B.1.3	Details of extensions	181
B.1.4	Details of the price setting application and simulation study	183
B.1.5	Details of the heterogeneous agent application	184
C	Appendix for Chapter 3	187
C.1	Notation and Least Squares for the Standard Linear Model	187
C.2	Bayesian Estimation of the Linear Model	192
C.2.1	Coefficient Estimation in Known- or Conditional-On- Σ Case	193

C.2.2	Estimation under Dependent Normal-IW Prior	197
C.2.3	Estimation under Independent Normal-IW	200
C.2.4	Distributions	201
C.3	Bayesian OVB Results	203
C.3.1	OV B Results under Prior of Lemma 13	203
C.3.2	OV B Results under Prior of Lemma 14	210
C.3.3	Results for Adjustment Procedures	214
C.4	Additional Prior Setting Info for BVARs	216
C.5	Empirical Application	218
C.5.1	Data	218
C.5.2	Additional Figures for Nowcasts	220
C.5.3	Additional Empirical Results	222

Chapter 1

Standard Errors under Correlated Assignment in Spatial Settings

1.1 Introduction

Causal inference under treatment effect heterogeneity typically assumes assignment to treatment that exhibits *independence* across units or non-overlapping clusters. Yet settings exist in which treatment status for individual units are correlated but units cannot be ex ante partitioned into non-overlapping clusters for standard cluster-based inference.

As described in [Borusyak and Hull \(2023\)](#), such “fuzzy clustering” naturally arises in network and spatial causal inference settings, the latter of which is the focus of this paper. In spatial settings, interventions occur at *specific* geographic locations, inducing treatment in surrounding units. As a consequence, the treatment status of geographically proximate units are typically correlated. Previous work demonstrates the necessity of adjusting standard errors in causal settings with correlated treatment and that clustering is sufficient when clusters are collectively jointly assigned (see [Abadie, Athey, Imbens and Wooldridge \(2017\)](#) and [Barrios, Diamond, Imbens and Kolesár \(2012\)](#)). However, unlike standard settings, in spatial applications, one typically *cannot* ex ante partition units into disjoint clusters that are always collectively assigned, which rules out standard clustering adjustments for inference. For example, the distribution of treatment locations may be continuously distributed over the spatial domain as in many urban and dense population settings ([Brown, 2018](#)). Here it is impossible to characterize the units as belonging to disjoint clusters

ex ante, although *treatment* typically still exhibits correlation or “fuzzy clustering.” Unlike the standard clustered assignment case, less concrete guidance exists for conducting valid causal inference about average treatment effects in this fuzzy clustering setting.

Further complicating matters, spatial causal inference settings with fuzzy clustering typically *also* involve pronounced nonstationarity in outcomes. For example, [Bera and Simlai \(2005\)](#) and [Banerjee, Gelfand, Knight and Sirmans \(2004\)](#) have argued for spatial nonstationarity in house prices, reflecting differences in the variance of house prices and correlation between house prices at a given distance over the spatial domain of study, even after adjusting for covariates. Such differences in the spatial distribution of potential outcomes and individual treatment effects could arise prior to treatment from agents’ sorting ([Couture, Gaubert, Handbury and Hurst, 2019](#)) or fixed geographic features of a given spatial domain ([Saiz, 2010](#)).

In response, this paper states a set of results for conducting valid causal inference in spatial binary treatment settings while being robust to both fuzzy clustering and spatial nonstationarity in potential outcomes. I show how knowledge of the assignment mechanism can be exploited to verify spatial weak dependence of treatment status, which is then used to construct an asymptotically normal estimator of the average treatment effect (ATE) and conservatively estimate its variance. While a strong assumption, knowledge of some features of the assignment mechanism is necessary in *any* causal inference setting to consistently estimate average treatment effects (ATEs) in the first place. This is especially true in spatial settings where geographic assignment typically leads to unequal treatment probabilities across units, which must be addressed to avoid bias in estimating the ATE ([Borusyak and Hull, 2023](#)). And so by further exploiting the assignment mechanism directly, one can correct for fuzzy clustering with the added benefit of *also* accommodating arbitrary heterogeneity in potential outcomes—e.g. from sorting, sharp changes at borders, etc.

In terms of implementation, the asymptotically normal ATE and variance estimators are constructed from (joint) treatment probabilities deduced from the assignment mechanism; inference is done using the corresponding t-statistic, where variation in this statistic stems from randomization over assignment. The construction of the ATE and variance estimators largely follows [Aronow, Samii et al. \(2017\)](#). Yet distinct from their approach, I establish the asymptotic behavior of the estimators under explicitly spatial limit theorems from [Jenish and Prucha \(2009\)](#), which exploit more traditional weak dependence assumptions on the binary treatment indicators rather than somewhat non-spatial independence conditions. This also makes clear that the assumption of binary treatment is less important for inference than the route from spatial weak dependence to valid design-based standard errors. Second, using the randomization distribution over assignment for inference means that inference is fundamentally *conditional* on potential outcomes, which

permits zero distributional assumptions on causal effects.

After stating the recommended inference approach and standard errors, I also show that these standard errors are preferable to those of Conley (1999) in terms of coverage and power for design-based inference. In the spatial econometrics literature, the Conley (1999) variance estimator was originally motivated as a spatial analog to HAC estimators from the time series literature, providing standard errors for GMM settings with spatially correlated *outcomes*. One might reasonably hope that Conley (1999) standard errors also enable appropriate design-based inference under correlated *treatment*. However, as I show, Conley (1999) standard errors typically lead to design-based inference that is excessively conservative under positive correlation in heterogeneous treatment effects (which is likely the most relevant case in practice), relative to the Aronow et al. (2017) estimator that fully exploits knowledge of the assignment mechanism.¹ This translates into excessively *wide* confidence intervals under positive correlation in individual treatment effects and power loss. Therefore, Conley (1999) standard errors are appropriate only for a superpopulation inference perspective (rather than design-based) or for guarding against interference, as recently discussed in Leung (2022) and Wang, Samii, Chang and Aronow (2023).

Following the asymptotic results, I use a simple analytical example and simulation study to compare variance estimators and how performance varies with correlation in treatment, the distribution of heterogeneous treatment effects, and sample size. As a first step, I demonstrate that the variance estimator of Aronow et al. (2017) succeeds in achieving nominal coverage in large samples, across different degrees of positive spatial correlation in treatment status, which is likely the most relevant case. I also document the large-sample power loss from using SHAC estimators in a design-based setting in this simple example. Lastly, the simulation studies also suggest that the variance estimator of Aronow et al. (2017) suffers from small-sample bias that leads to undercoverage. This suggests bias adjustment in variance estimation may also be an important area of future research in spatial design-based settings with correlated treatment assignment, similar to the important literature on finite-sample HAC variance estimation (Lazarus, Lewis, Stock and Watson, 2018).²

Finally, in an instrumental variables (IV) setting with known assignment mechanism for the instrument (that allows correlated treatment across units), I show how the variance estimator construction of Aronow et al. (2017) can be repurposed to construct a variance estimator for a LATE estimand under general

¹The conclusions of the comparison will hold for other SHAC estimators (which essentially differ only in finite sample properties) or robust alternatives like Müller and Watson (2022). The ranking of variances results from the difference in inference perspectives (design-based vs. superpopulation) and assumptions placed on potential outcomes (none vs. stationary)

²Qian (2020) is a recent example from the survey literature, which discusses finite-sample corrections in cases of sampling without replacement. Since design-based causal inference is deeply related (since one samples units for population and control separately without replacement), the results could likely be extended in a straightforward way. In the econometrics literature, Hong, Leung and Li (2019) have studied the effect of limited overlap on bias, whereas I instead document how the data generating process for treatment effects seems to play a role, separate from overlap.

treatment effect heterogeneity. Instrumental variables is quite common in spatial and network settings (see [Baum-Snow and Ferreira \(2015\)](#) and [Borusyak and Hull \(2023\)](#) for examples), as spatial and geographic settings in economics bring with them substantial endogeneity issues, even setting aside general equilibrium issues. This result therefore enables IV inference under treatment effect heterogeneity that again properly accounts for “fuzzy” clustering in treatment assignment. Finally, because the IV variance estimator takes the form of the [Aronow et al. \(2017\)](#) estimator from the non-IV case, the intuition and results developed in the non-IV setting carry over directly to the IV setting.

Preceding Literature While the causal econometric literature on spatial assignment is fairly recent, there has been previous applied work on estimating the effects of spatial treatment. Such place-based treatment may include the location of power plants ([Davis, 2011](#)), grocery stores ([Freedman and Kuhns, 2018](#)), foreclosures ([Ellen, Lacoë and Sharygin, 2013](#)), low-income housing ([Baum-Snow and Marion \(2009\)](#) and [Diamond and McQuade \(2019\)](#)), or public health interventions ([Miguel and Kremer, 2004](#)). These interventions then jointly affect nearby house prices, crime, public health, etc. in the vicinity of the intervention. See [Baum-Snow and Ferreira \(2015\)](#), [Borusyak and Hull \(2023\)](#), and [Pollmann \(2023\)](#) for further examples. The proliferation of spatial data has increased the demand for econometric guidance.

Until recently, [Conley \(1999\)](#) standard errors have been the standard for inference in spatial settings (when clustering at the level of assignment is not possible), even though this paper shows they are likely to be too conservative for ATE estimation under heterogeneous treatment effects and correlated treatment. The [Conley \(1999\)](#) variance estimator is a special case of the broader class of spatial heteroskedasticity and autocorrelation (SHAC) standard errors, which are a two-dimensional analog to time series HAC estimators from the time series literature. They were originally motivated under spatial stationarity assumptions on outcomes, which is ill-suited to causal inference with treatment effect heterogeneity.³ Moreover, their performance and properties in a potential outcomes framework with a known assignment mechanism has only recently been explored by this paper and contemporaneous papers in the literature.

Instead, this paper will follow the broader causal literature and make assumptions on assignment so as to avoid sampling or modeling assumptions on outcomes. This direction follows papers like [Abadie, Athey, Imbens and Wooldridge \(2020\)](#), which studied regression settings with *independent* assignment. Instead, I study *joint* assignment with binary treatment, although results could be extended to the multivalued treatment case at the cost of additional notational complexity and requirements on treatment probabilities.⁴

³As in the time series case ([Andrews, 1991](#)), stationarity is not strictly needed for SHAC estimation, just weak dependence on outcomes. However, even weak dependence for outcomes may be undesirable given strong spatial sorting.

⁴The results of [Abadie et al. \(2020\)](#) in the case of single binary treatment regression are nested by the results in this paper.

This design-based perspective is relevant not only because it will relax assumptions on outcomes, but in some experimental settings, researchers may even control the spatial assignment mechanism directly.⁵

Many of the results in [Abadie et al. \(2020\)](#) have their counterpart in this paper with joint assignment; however, the poor performance of [Conley \(1999\)](#) standard errors stems from *correlation* in treatment, which is particular to this paper and other papers that consider joint as opposed to individualistic and independent assignment; see the next section of the literature review. On the similar side, [Abadie et al. \(2020\)](#) showed that that Eicker-Huber-White (EHW) standard errors are typically larger than those implied by purely design-based randomization, although one cannot reduce the bias since it depends upon unobserved (heterogeneous) treatment effects. This paper similarly finds that [Conley \(1999\)](#) and SHAC standard errors are typically larger than those implied by purely design-based randomization. However, I demonstrate that one *can* do better than SHAC variance estimators by fully exploiting information about joint treatment probabilities from the known assignment mechanism.

Finally, to formalize a random process for spatial assignment, I draw briefly on the spatial statistics literature for point processes. Most papers in the spatial literature drastically simplify the assignment mechanism, e.g. designating a fixed number of “potential sites” and independently assigning each site to treatment or not based on a Bernoulli trial ([Wang et al., 2023](#)). However, I discuss and show in a simple example how a spatial point process—a common construction in spatial statistics ([Diggle, 2013](#))—can be used as a tool to formalize any spatial random assignment mechanism, more generally.

Recent and Contemporaneous Literature This paper is closely related to recent papers in the causal econometric literature that all use “design-based” inference to address the shortcomings mentioned above. This focus on design-based inference naturally follows because the spatial literature is a particular instance of a “bipartite” causal inference problem ([Zigler and Papadogeorgou, 2021](#)), in which the level of assignment is distinct from the level at which outcomes are measured.⁶ As emphasized in this paper, [Borusyak and Hull \(2023\)](#), and [Pollmann \(2023\)](#), a design-based (rather than random sampling) perspective is preferable when assignment and outcomes exist at different levels. Recent papers differ in their assumptions on assignment, potential outcomes, and (as a consequence) their target estimands, which I now discuss.

The first related paper is [Aronow et al. \(2017\)](#), (1) that I use to construct a variance estimator with better properties [Conley \(1999\)](#) under spatial correlation in treatment and (2) that I extend to the instrumental

⁵In [Blattman, Green, Ortega and Tobón \(2017\)](#), the authors could choose where to randomly assign police officers or neighborhood improvements. In [Mobarak and Rosenzweig \(2013\)](#), researchers randomly located rainfall measurement stations to measure the effect of distance to measurement stations on insurance uptake when payouts are based on rainfall measured at that precise station location.

⁶For example, treatment is assigned to specific locations in space, not paired with individual units.

variables case, which is relevant for applications. While the framework of [Aronow et al. \(2017\)](#) is general and uses a CLT that relies on limited network dependency for treatment, I adapt their ATE and variance estimator to the spatial case and instead demonstrate asymptotic normality of the ATE estimator under spatial weak dependence for treatment.⁷ I then make a natural comparison to the [Conley \(1999\)](#) variance estimator, which is a new contribution not considered in the setting with correlated treatment.⁸ I also demonstrate that this variance estimator is preferable to [Conley \(1999\)](#) standard errors which are otherwise typically too conservative. Finally, I show that extension to LATE estimators is straightforward, and one can use a modification of their variance estimator to derive standard errors for a LATE estimand. Given the same spatial weak dependence assumptions on assignment for an instrument, asymptotic normality holds for LATE-related quantities. This extension was missing from the literature, including the next paper.

Next, within the econometric literature, [Borusyak and Hull \(2023\)](#) have recently emphasized the importance of adjusting effect estimators for differential treatment probabilities to avoid bias, while my paper focuses on inference and variance estimation beyond the cases considered by [Borusyak and Hull \(2023\)](#). In particular, they consider a regression IV setting with continuous regressors and constant marginal effects, and “debias” estimators of the effects by including an expected instrument within the regression. In this paper, I focus on a binary treatment setting (for simplicity) with *heterogeneous* individual treatment effects, and account for differential treatment probability by inverse-probability weighting.⁹ While they focus on bias and test only sharp null hypotheses using Fisher randomization inference, I focus on asymptotically normal inference under treatment effect heterogeneity. I do this not only for the ATE, but also a LATE estimand under treatment effect heterogeneity. And while the authors do show that their estimation approach recovers a LATE estimand in the case of heterogeneous treatment effects, they do not prove asymptotic normality or provide a corresponding variance estimator. In this paper, I show how one can translate weak dependence in spatially correlated treatment into an asymptotically normal estimator of the LATE (in the binary treatment case), and state a variance estimator. Importantly, to derive this variance estimator, I do not need to impose more on the problem than [Borusyak and Hull \(2023\)](#). They already assume that one can simulate from the assignment mechanism, so my results simply use additional moments generated by the already assumed assignment mechanism to accommodate heterogeneity.

Distinct from [Aronow et al. \(2017\)](#), [Borusyak and Hull \(2023\)](#), and this paper, there is another strand of the recent literature that studies interference, which I do not consider. Specifically, [Pollmann \(2023\)](#)

⁷CLTs under spatial weak dependence are already familiar from the SHAC literature, e.g. [Jenish and Prucha \(2009\)](#). The only difference is that I apply the limited dependence assumption treatment rather than outcomes themselves.

⁸See subsequent discussion on using [Conley \(1999\)](#) for interference and network dependence, which is a distinct use case.

⁹Under binary treatment, IPW and the expected instrument approach are exactly equivalent.

and Wang et al. (2023) allow for interference in which an individual/unit’s outcome depends upon the full realization of spatial assignment, rather than an individualistic treatment status.¹⁰ In other words, they assume that one must *always* write potential outcomes for any unit i as $Y_i(\{s_c\}_{c=1}^C)$ where $\{s_c\}_{c=1}^C$ is the full list of treatment locations, and that one cannot simplify or reparametrize this mapping to $Y_i(D_i)$, where D_i is an individualistic “exposure,” “sufficient statistic” for treatment, or instrumental variable associated with unit i . I assume the latter, while they do not. Because they allow potential outcomes to depend upon multiple nearby treatment locations and interact in arbitrarily complex ways, their estimands must marginalize over the assignment mechanism and, as a consequence, depend upon it.¹¹

Pollmann (2023), therefore, differs mainly in the target estimand, the assumptions on assignment, and in his novel extension to settings with unconfoundedness. The estimand of Pollmann (2023) is the average treatment effect on the treated (ATT), allowing for general interference, while I rule out interference and focus on the ATE and LATE. To simplify, the author also assumes that treatment is assigned to separate regions or independently to pre-selected candidate locations. This choice is driven by the desire to handle interference, while my stronger assumptions on potential outcomes (e.g. spatial treatment manifesting fully as a binary “exposure” D_i for each unit, rather than dependence on the full configuration $\{s_c\}_{c=1}^C$) allows me to adopt weaker assumptions on the spatial assignment mechanism. In general, this tradeoff is common. General interference often requires simpler designs for more interpretable results. In contrast, my stronger assumptions on potential outcomes allow weaker assumptions on assignment. Finally, Pollmann (2023) offers a useful practical approach to matching treatment and control locations under unconfoundedness, which I do not consider.

Similar to Pollmann (2023), Wang et al. (2023) also study design-based inference for spatial assignment under interference, with a difference essentially only in the estimand and how it weights individual units.¹² Therefore, all comments for Pollmann (2023) carry over.

Finally, Wang et al. (2023) is also notable, along with Leung (2022) from the network causal inference literature, because they also evaluate the performance of the Conley (1999) variance estimator. Leung (2022) uses the Conley (1999) variance estimator to address interference in a network causal inference setting with

¹⁰This manner of interference is from (1) spatial assignment spilling outward over space and (2) individuals’ outcomes depending upon multiple treatment locations in complex ways. This manner of interference is distinct from units’ outcomes depending directly upon outcomes of other nearby units, which is the network setting as in Leung (2022). See Pollmann (2023) for discussion.

¹¹That is, unless one is willing to make stronger modeling assumptions or exclusion restrictions on potential outcomes.

¹²While Pollmann (2023) conducts inference treating individuals as the unit of observation either assigned to treatment or not, Wang et al. (2023) redefines the unit of observation to be circle averages around potential treatment sites, which are either assigned to treatment or not. However, it is also worth noting that Pollmann (2023) also states results for this case as well, although that is not his leading estimand.

independent assignment. Here, [Conley \(1999\)](#) standard errors help guard against unmodeled spillovers that exist even despite correlated assignment. In contrast, my causal inference problem does not have spillovers between units, but rather in spatial assignment/exposures that induces correlation in treatment. After accounting correlation in treatment, there is no further justification for [Conley \(1999\)](#) standard errors, and in fact, they are typically too conservative because (based on the results of [Leung \(2022\)](#)) they are additionally accounting for spillovers. With known assignment mechanism and no interference—but through spatial proximity to treatment as in my paper and [Borusyak and Hull \(2023\)](#)—one can do better than [Conley \(1999\)](#) standard errors by exploiting knowledge of joint treatment probabilities, which is only relevant with correlated assignment. Finally, [Leung \(2022\)](#) also compares the variance estimator of [Aronow et al. \(2017\)](#) to the [Conley \(1999\)](#) under “correct specification” of the exposure mapping, which is what I consider. However, I modify the variance estimator slightly to more directly compare it to the [Conley \(1999\)](#) variance estimator. This modification is important, because without it, [Leung \(2022\)](#) cannot rank [Conley \(1999\)](#) and a [Aronow et al. \(2017\)](#)-based estimator, while I can.

Lastly, although [Wang et al. \(2023\)](#) is motivated as a spatial assignment problem, the assumptions placed on assignment and outcomes mean that it is a network example, and hence a special case of [Leung \(2022\)](#). As the authors state, the spatial problem they consider is equivalent to a network causal inference problem in which potential treatment sites are independently treated or not, and there are unmodeled network linkages between potential sites.

Outline Section 2 sets up the causal and asymptotic framework for analysis, as well as an analytical example that will be used to provide intuition and simulation results. Having set up the causal framework, Section 3 states the effect estimator and proves asymptotic normality under spatial weak dependence. Section 4 contains the main results on variance estimation, including a comparison to [Conley \(1999\)](#) standard errors and extended discussion on its behavior relative to the preferred variance estimator. Having fully described an approach to inference—from potential outcomes to asymptotic normality and variance estimation—Section 5 applies all results to the simple analytical example and offers a small simulation study. Finally, Section 6 extends the results of the previous sections from the ATE to cover the LATE case.

1.2 Framework and Preliminaries

This section introduces notation and a framework for spatial causal inference problems in order to precisely state subsequent estimators and asymptotic results for them. Like any causal inference problem, one must

describe the random assignment mechanism, potential outcome functions given realized treatment assignment, and the population estimands of interest. At the end of this section, I also introduce an analytical example that will be used for illustrating results and the subsequent simulation study.

1.2.1 Spatial Assignment Mechanism

Treatment is characterized by a *spatial point process* over a two-dimensional spatial domain $S \subset \mathbb{R}^2$. A realization from this point process implies an (assumed) finite number of treatment centers C , placed at specific locations

$$\{s_c\}_{c=1}^C \subset S \subset \mathbb{R}^2.$$

In some cases, e.g. Blattman et al. (2017), the assignment mechanism may be controlled by the researcher. In other cases, the assignment mechanism might be assumed by the researcher given knowledge of how a particular policy rule determined treatment center locations.

This characterization of the assignment mechanisms illustrates the bipartite nature of the causal inference problem (Zigler and Papadogeorgou, 2021). As discussed in Pollmann (2023) and apparent in the above description, the process governing assignment is wholly separate from individual units. Treatment is not individualistic or perfectly isolated and paired with particular units, which will drive fuzzy clustering. Rather, it occurs at the level of the spatial domain.

Remark. (Generality of Spatial Point Processes) Spatial point processes are the most general way of encoding *any* spatial assignment mechanism; therefore, they can be used as a general construct for capturing one part of the bipartite spatial causal inference problem. For example, spatial point processes nest random assignment of a fixed number of centers to preselected “potential sites”, as in Davis (2011) or Wang et al. (2023).¹³ They also include Poisson processes, in which a density or intensity function over the domain S encodes the probability of treatment in a given subset of S .¹⁴ See the analytical example in Section 1.2.5 for a simple Poisson point process assignment mechanism.

Remark. (Role of Weak Dependence) The literature on spatial point processes is deep and somewhat technical,¹⁵ but ultimately, what matters is whether there is enough weak dependence in the spatial assignment mechanism to, in turn, induce enough weak dependence in outcomes, which enables inference. Said another

¹³See Borusyak and Hull (2023) for further examples.

¹⁴See Appendix A.1 for the formal definition.

¹⁵See Diggle (2013) for a textbook treatment of spatial point process theory.

way, the asymptotic sequence of point processes should imply that treatment in one region is not excessively correlated with treatment in another region. What exactly is “enough” weak dependence (and how to measure it) will depend also on the potential outcomes function, where I discuss this concept further.

1.2.2 Individual Units and Potential Outcomes

Population Within the spatial domain $S \subset \mathbb{R}^2$ (to which treatment is assigned), there exists a finite population of N total units residing at locations ℓ_i on a lattice $\mathcal{I}$:

$$\{\ell_i\}_{i=1}^N \subset \mathcal{I} \subset S$$

Throughout, I assume throughout that the researcher has access to the full finite population of units, rather than some sample. This choice is purely for convenience and to more clearly exposit the inferential approach with less overhead. One could extend the results to allow for sampling on top of randomized assignment. See [Abadie et al. \(2020\)](#) for an example that includes both forms of randomness.

Potential Outcomes With treatment and outcomes occurring at different levels in the bipartite spatial setting, potential outcomes for unit i could depend upon the full realization of spatial treatment. Therefore, to capture the most general case, potential outcomes must be written

$$Y_i(\{s_c\}_{c=1}^C).$$

This generality means that inference must use either (1) nonstandard estimands that require marginalization to handle arbitrary spillovers and interference or (2) parametric modeling assumptions on the potential outcome functions. See [Pollmann \(2023\)](#) and [Wang et al. \(2023\)](#) for examples in (1), which is one recent approach the literature has taken.

Instead, I will make simplifying assumptions so that I can draw a more precise comparison between variance estimators in the fuzzy clustering case. I assume throughout this paper that treatment for units can be summarized by an individualistic binary treatment indicator

$$D_i = D_i(\ell_i, \{s_c\}_{c=1}^C) \tag{1.1}$$

so that the general potential outcomes function can be reparameterized simply as

$$Y_i(\{s_c\}_{c=1}^C) = Y_i(D_i). \quad (1.2)$$

This imposes the usual stable unit treatment value assumption (SUTVA), in that outcomes do not depend upon the outcomes of other units directly but only an individualized treatment indicator (which will still be correlated across units).

Remark. (Exposure Mappings) Expression 1.1 is often called an “exposure mapping”.¹⁶ It maps unit i ’s location and the underlying random spatial assignment process with realization $\{s_c\}_{c=1}^C$ into exposures for individual units, according to some rule.

Remark. (Example Exposure Mapping) A simple example for the treatment rule (or exposure mapping) would be a case where binary treatment is determined by whether or not a treatment center is within some cutoff distance of the unit.

$$D_i = \mathbf{1}\{d(\ell_i, \{s_c\}_{c=1}^C) \leq \delta\} \quad (1.3)$$

where $d(\ell, \{s_c\}_{c=1}^C)$ gives the Euclidean distance between location ℓ and the nearest treatment center in the set of treatment centers $\{s_c\}_{c=1}^C$, and where δ is some fixed known cutoff. In this way, randomly placed treatment centers treat nearby units, inducing fuzzy clustering.

Treatment Probabilities The simplifying assumptions on potential outcome functions, summarized in Expressions 1.1 and 1.2, imply that the researcher can compute

$$\pi_i = P[D_i = 1] \quad \pi_{ij} = P[D_i = 1, D_j = 1]$$

In other words, the mapping in Expression 1.1 plus the random assignment mechanism for treatment locations induces a joint distribution for the treatment indicators $\{D_i\}_{i=1}^N$. Because of “fuzzy” clustering, treatment indicators are not all independent, and so usually $\pi_{ij} \neq \pi_i \pi_j$.

In this paper, these treatment probabilities are *all* that’s needed from the spatial assignment mechanism to construct estimators. More generally, for bipartite spatial causal inference problems, treatment probabilities

¹⁶See Aronow et al. (2017) in the causal literature, Leung (2022) in the network literature, and Borusyak and Hull (2023) in the econometric literature.

are sufficient statistics of the full spatial assignment mechanism when it comes to constructing estimators.¹⁷ And while these probabilities are often not analytically available, they can be approximated to arbitrary precision by simulating from the known spatial assignment mechanism, applying the treatment rules in Expression 1.1, and averaging across simulations.

1.2.3 Estimand

Given the assumptions I have placed on potential outcomes, I focus on estimation of the finite sample average treatment effect (ATE),

$$\alpha_{ATE} = \frac{1}{N} \sum_{i=1}^N Y_i(1) - \frac{1}{N} \sum_{i=1}^N Y_i(0).$$

Under the binary treatment regime described above, the ATE compares average outcomes under the hypothetical scenarios in which *all* units are affected by a nearby treatment center versus not. Given a treatment rule like Equation (1.3), this estimand is appropriate for policy makers deciding whether or not to densely pack treatment centers within the domain to “cover” all units.

Moreover, even if the ATE is not ultimately the estimand of interest, it represents a useful starting point for describing the estimation and inference strategy that can be further modified and adapted in straightforward ways. As a first example, suppose the assignment mechanism is such that certain subregions of S or types of neighborhoods may never be subjected to treatment. In such cases, the estimand may instead be a *conditional* average treatment effect (CATE) that conditions on membership in the relevant subregions or types of neighborhoods. The researcher need only ensure that the specified assignment mechanism reflects this and take care to interpret the estimand as a CATE. The estimation and inference steps described below will otherwise look the same. Second, even if the treatment rule in Expression 1.1 is non-binary, one can define contrasts of different treatment levels as estimands.

1.2.4 Asymptotic Embedding and Inference Perspective

This section defines the asymptotic embedding that underpins inference. I also stress the usefulness of the design-based perspective inherent to this embedding, which permits inference under weaker assumptions.

The ingredients to the problem for any finite population are thus the domain S , the lattice of locations for the N units $\{\ell_i\}_{i=1}^N = \mathcal{I}$, the treatment indicators $\{D_i\}_{i=1}^N$, the assignment mechanism that induces

¹⁷That said, more information about the assignment mechanism is necessary to ensure the weak dependence needed for asymptotic normality of the ATE estimator.

randomness in treatment (along with the rule), corresponding individual and pairwise treatment probabilities, $\{\pi_i\}_{i=1}^N$ and $\{\pi_{ij}\}_{i,j=1}^N$, and finally potential outcomes $\{Y_i(0), Y_i(1)\}_{i=1}^N$.

To establish asymptotic results, all objects in a given finite population are recast as elements in a triangular *sequence* of growing populations with increasingly many units spread over increasingly large domains,

$$S_n, \quad \{\ell_{n,i}\}_{i=1}^{N_n} = \mathcal{I}_n, \quad \{D_{n,i}\}_{i=1}^{N_n}, \quad \{\pi_{n,i}\}_{i=1}^{N_n}, \{\pi_{n,ij}\}_{i,j=1}^{N_n}, \quad \{Y_{n,i}(0), Y_{n,i}(1)\}_{i=1}^{N_n},$$

with the subscript n indexes a particular spatial random experiment within the sequence of experiments.¹⁸

This paper’s “design-based” inference perspective means that each experiment in the sequence is random only due to the assignment mechanism only, rather than random sampling; therefore, I do not appeal to an artificial superpopulation. Inference, for any n , will be conditional on potential outcomes, the assignment mechanism, and the spatial distribution of units. Weak dependence of spatial assignment and, as a consequence, treatment indicators will deliver asymptotic normality, rather than weakly dependent outcomes drawn from a hypothetical superpopulation.

For spatial settings, this design-based perspective is desirable and often necessary, as it allows inference under weaker assumptions. Heterogeneous units and individuals are rarely scattered randomly across the spatial domain,¹⁹ contiguous regions often exhibit unknown dependence due to sorting, and regions exhibit considerable observable and unobservable heterogeneity that could affect outcomes. Therefore, possessing a truly random, representative sample from the superpopulation is doubtful. This paper and the design-based spatial literature offer a useful alternative for inference, similar to Fisher randomization inference, which has already been employed in practice in applications to test sharp null hypotheses. See [Blattman et al. \(2017\)](#), [Borusyak and Hull \(2023\)](#), [Pollmann \(2023\)](#) for examples.

Also, validity is not lost by adopting this asymptotic and inference perspective. If potential outcomes are, in fact, drawn from some spatially stationary superpopulation distribution, inference and confidence intervals valid under the design-perspective *conditional* on potential outcomes will also be valid *unconditionally*. The estimand $\alpha_{n,ATE}$ will converge to the superpopulation ATE, and the estimator $\hat{\alpha}_{n,ATE}$ concentrated around $\alpha_{n,ATE}$ (by design-based consistency) will then also concentrate around the superpopulation ATE. Second, if a researcher only has access to a sample from the population (rather than the full population as assumed), a sampling mechanism can be added as an additional ingredient within the triangular sequence of experiments. A researcher could thus incorporate sampling uncertainty, while maintaining the attractively

¹⁸In discussion, I will sometimes drop the subscript n . For regularity assumptions and proofs, I will retain the subscript.

¹⁹See [Pollmann \(2023\)](#) for a discussion of the problems for inference that arise when the potential outcome distribution depends upon location, rather than being randomly distributed.

weak assumptions on the cross-sectional and spatial distribution of individual treatment effects stressed in this subsection. See [Abadie et al. \(2020\)](#) for the case where treatment is individually assigned.

1.2.5 Analytical Example

In this section, I outline all pieces of an analytically tractable spatial causal inference problem, which will be useful for demonstrating subsequent results. Conveniently, the degree of fuzzy clustering will be indexed precisely by the relationship between a few key parameters in this simple problem.

Definition 1. (Domain and Individual Units) The domain S is a circle with circumference $|S| = N \cdot \Delta$, for some fixed constant $\Delta > 0$. On this circle live N *evenly spaced* units at locations

$$\{\ell_i\}_{i=1}^N = \{(i-1)\Delta\}_{i=1}^N \subset S,$$

where location $\ell = 0$ is chosen to be the top of the circle. Hence, Δ is the arc length between adjacent units.

Definition 2. (Spatial Assignment Mechanism) Treatment centers are assigned to locations on S according to a homogeneous Poisson point process. Therefore, the number of treatment centers C is a $\text{Poisson}(\lambda|S|)$ random variable, for some intensity λ , and the C treatment centers are independently and uniformly scattered around the circular domain. See [Appendix A.1](#) for a formal definition.

The above assignment mechanism corresponds to the simplest point process and illustrates how point process theory from the spatial statistics literature could be used to formalize spatial random assignment. There also exist straightforward and attractive extensions like *inhomogeneous* Poisson point processes in which a nonnegative intensity function $\lambda(s)$ varies over points $s \in S$ (perhaps as a parametric function of region-specific covariates) with higher values of the intensity $\lambda(s)$ at a given point s encoding a higher probability of assignment within a small neighborhood of that point.

Definition 3. (Treatment Rule) Treatment follows Expression [1.3](#), i.e.

$$D_i = \mathbf{1}\{d(\ell_i, \{s_c\}_{c=1}^C) \leq \delta\}$$

Proposition 1. *The circular domain, the configuration of units, the assignment mechanism, and the treatment*

rule together together imply treatment probabilities

$$\begin{aligned}\pi_i &= P[D_i = 1] = 1 - e^{-2\delta\lambda} \\ \pi_{ij} &= P[D_i = 1, D_j = 1] = 1 - 2e^{-2\delta\lambda} + e^{-\lambda[2\delta + \min\{2\delta, d(\ell_i, \ell_j)\}]} \\ &= \begin{cases} 1 - 2e^{-2\delta\lambda} + e^{-\lambda[2\delta + \min\{2\delta, d(\ell_i, \ell_j)\}]} & d(\ell_i, \ell_j) \leq 2\delta \\ \pi_i \pi_j & d(\ell_i, \ell_j) > 2\delta \end{cases}\end{aligned}$$

where $d(\ell, \ell')$ is the arc length between locations ℓ and ℓ' on the circle.

Proof. See [Appendix A.1.2](#). □

Remark. Notice that all individuals have identical first-order treatment probabilities without any adjustments needed at some boundary or edge. This motivated the choice of a circular domain with simple homogeneous Poisson (uniform) assignment.

As can be seen in the probabilities above, the degree of fuzzy clustering is characterized by two key parameters, which proves useful for building intuition and demonstrating results below. First, increasing λ directly increases the expected number of treatment centers and corresponding first order treatment probability. Second, parameter δ controls not only treatment probability, but also the degree of fuzzy clustering that is due to the treatment rule and spacing of units, holding fixed the assignment mechanism. Notice from the expression for π_{ij} that treatment indicators D_i and D_j are independent beyond the distance 2δ —precisely when units cannot be treated by the same treatment center and therefore never fall into a fuzzy cluster. Therefore, not only do larger values of δ increase the probability a given unit will be treated, but also correlation in treatment status.

Finally, this simple setup also permits straightforward asymptotics. For asymptotics, I will hold fixed λ , δ , and Δ , and simply increase N , which will scale the size of the circle as well as the number of units, while holding fixed the proximity of units, intensity of treatment, and degree of fuzzy clustering as a consequence. Notice also that in this asymptotic framework, there are increasingly many units, but they are not getting closer together. This is the sense in which the asymptotics are of “expanding domain” type, rather than “infill” type.

1.3 Estimation of the ATE

This section defines the ATE estimator in this spatial setting, proves consistency under weak conditions, and proves asymptotic normality under an explicitly spatial central limit theorem. With the exception of assuming bounded potential outcomes, all results stem from regularity conditions (e.g. weak dependence) placed on the triangular sequence of binary treatment indicators $\{D_{n,i}\}_{i=1}^{N_n}$ only. No other explicit restrictions are placed on the spatial domain, units' locations, assignment, or treatment rules except insofar as they affect the sequence of joint distributions for the treatment indicators.

The next section, which contains the main results on variance estimation, builds on the ATE estimation results of this section and will use similar weak dependence conditions to (1) prove consistent estimation of the ATE estimator variance and (2) compare the proposed variance estimator to spatial HAC.

1.3.1 Effect Estimator

In this section, I define the ATE estimator, which is fixed despite the approach one ultimately chooses for variance estimation. Despite the complications for variance estimation, estimation of the ATE is straightforward. The estimator is simply a weighted least squares (WLS) estimator, weighting each observed outcome by the inverse probability of observed treatment, deduced from the assignment mechanism.

Definition 4. (WLS Estimator) Let $\widehat{\beta}_{n,WLS}$ denote the estimator that results from weighted least squares regression of observed $Y_{n,i}$ on $X_{n,i} = (1 - D_{n,i}, D_{n,i})$ using the following weights

$$\omega_{n,i} = \frac{D_{n,i}}{\pi_{n,i}} + \frac{1 - D_{n,i}}{1 - \pi_{n,i}}$$

constructed from observed treatment status and individual treatment probabilities. Also let

$$\widehat{\alpha}_{n,ATE} := \begin{pmatrix} -1 & 1 \end{pmatrix} \widehat{\beta}_{n,WLS}.$$

Remark. (Importance of Weighting) Notice that ordinary least squares regression of observed outcomes $Y_{n,i}$ on a constant and a treatment indicator $X_{n,i} = (1, D_{n,i})$ does not consistently estimate the ATE when treatment probabilities $\pi_{n,i}$ vary across units. Weighting is necessary. This reinforces the result of [Borusyak and Hull \(2023\)](#), who show that both instrumental variables regression (including ordinary least squares regression as a special case) typically fails to recover the causal effect of interest under complicated assignment

mechanisms in which the distributions of the randomly assigned instrument varies across units. They adjust by including an “expected instrument” in the regression (averaging the instrument over repeated assignment), while I adjust by inverse-probability weighting. In the case of a binary instrument, their expected instrument is the treatment probability and the estimators are identical.

Definition 5. (Horvitz-Thompson Estimators) Define the following Horvitz-Thompson estimators which weight by the inverse probability of treatment status

$$\begin{aligned}\hat{\mu}_{n,Y(1)} &= \frac{1}{N_n} \sum_{i=1}^{N_n} \frac{D_{n,i}}{\pi_{n,i}} Y_{n,i} & \hat{R}_{n,1} &= \frac{1}{N_n} \sum_{i=1}^{N_n} \frac{D_{n,i}}{\pi_{n,i}} \\ \hat{\mu}_{n,Y(0)} &= \frac{1}{N_n} \sum_{i=1}^{N_n} \frac{1 - D_{n,i}}{1 - \pi_{n,i}} Y_{n,i} & \hat{R}_{n,0} &= \frac{1}{N_n} \sum_{i=1}^{N_n} \frac{1 - D_{n,i}}{1 - \pi_{n,i}}\end{aligned}$$

Lemma 1. *Provided neither $\hat{R}_{n,1}$ nor $\hat{R}_{n,0}$ equal zero, the above-defined WLS estimator $\hat{\alpha}_{n,ATE}$ can also be written in terms of the Horvitz-Thompson estimators,*

$$\hat{\alpha}_{n,ATE} = \frac{\hat{\mu}_{n,Y(1)}}{\hat{R}_{n,1}} - \frac{\hat{\mu}_{n,Y(0)}}{\hat{R}_{n,0}}.$$

Proof. See [Appendix A.3.2](#). □

Remark. (Hajek vs. Horvitz-Thompson) The WLS estimator is equivalent to the Hajek estimator, and if the $\hat{R}_{n,1}$ and $\hat{R}_{n,0}$ were instead replaced by their expectation of 1, the resulting estimator of average causal effects would be the Horvitz-Thompson estimator. While both are consistent, inverse-probability weighted estimators, the WLS/Hajek estimator is preferred.

First, using the WLS regression estimator, I exactly what researchers using [Conley \(1999\)](#) standard errors would do for ATE estimation. Therefore, this fixes the ATE estimator, so that the only difference for inference comes in variance estimation.

Second, formal results from [Hahn \(1998\)](#) and [Hirano, Imbens and Ridder \(2003\)](#) show that the Hajek estimator is preferable on efficiency grounds (even though slightly biased) in the case where observations are a random sample from a superpopulation. Though my setting with fuzzy clustering and design-based inference is different, simulation evidence suggests that a similar pattern holds in practice, with division by $\hat{R}_{n,d}$ stabilizing testimations of corresponding $\hat{\mu}_{n,Y(d)}$. In addition, other papers in the literature have similarly argued in favor of the Hajek estimator, e.g See [Aronow et al. \(2017\)](#) and [Leung \(2022\)](#).

1.3.2 Consistency

To obtain consistency of the ATE estimator, I restrict potential outcomes to be nonexplosive, ensure all units can be treated by assignment, and ensure that correlation in assignment is not too severe. The first two assumptions are carried through to asymptotic normality as they are important conditions for estimation regardless of the assumptions about dependence or fuzzy clustering. The last condition restricting dependence is at its weakest here in the case of consistency, and will be implied by more stringent conditions placed later that ensure asymptotic normality.

Assumption 1. (*Bounded Potential Outcomes*) There exists some constant $M \in (0, \infty)$, such that the sequence of potential outcomes $\{Y_{n,i}(1), Y_{n,i}(0)\}_{i=1}^{N_n}$ satisfy the following bound:

$$\forall n \quad |Y_{n,i}(d)| \leq M \quad d \in \{0, 1\} \quad i \in \{1, \dots, N_n\}.$$

This is the only assumption on potential outcomes needed for inference.²⁰ All other substantive assumptions that justify inference rely on assumptions on assignment.

This assumption could be weakened, and the proofs of consistency and asymptotic normality will implicitly make clear where and how. However, I maintain this assumption throughout because it is standard in empirical applications and weakening would rely specifying more about how exactly the assignment mechanism and joint distribution of treatment interacts with the set of potential outcomes. I therefore avoid this unnecessary and practically unimportant extension.

Assumption 2. (*Nondegenerate Treatment Probabilities*) For some constant $\varepsilon > 0$ that is independent of n and unit index i , the distribution of treatment indicators $\{D_{n,i}\}_{i=1}^{N_n}$ satisfy

$$P[D_{n,i} = 1] = \pi_{n,i} \in [\varepsilon, 1 - \varepsilon] \quad \forall n, i. \quad (1.4)$$

For any finite sample, this ensures that all units have some nondegenerate probability of being in the treated or control group over repeated runs of the experiment; therefore, the ATE can be estimated in any finite sample by averaging over the assignment mechanism. For the asymptotics, because $\varepsilon > 0$ does not depend upon n , Expression 1.4 also ensures that the probability of treatment for any unit does not

²⁰In making this statement, I am excluding the substantive assumptions made on the potential outcome function that reduced the mapping from general dependence on the full treatment realization $\{s_c\}_{c=1}^C$ to a single sufficient, individualistic exposure variable D_i . However, the broader point is that design-based inference and weak dependence on treatment can be used to weaken parametric modeling assumptions on outcomes. Even under interference, this is the case (Pollmann (2023)).

trend towards degenerate as the sample grows. Rather than explicitly, this paper only implicitly restricts the spatial domain, unit locations $\{\ell_i\}_{i=1}^N$, spatial assignment mechanism, and treatment rule to ensure the above condition is satisfied. The researcher need only verify that however these pieces of the problem interact, all units have nondegenerate treatment probability.

Finally, even if this assumption is violated for the full population, the results and conclusions of this paper can still hold after restricting the population and reinterpreting the estimand. Specifically, if there are regions that are never treated, the researcher may restrict the population and focus on those regions in which [Equation \(1.4\)](#) holds, redefining the estimand accordingly to represent the conditional average treatment effect (CATE), as discussed in [Section 1.2.1](#). Using only those observations for which [Equation \(1.4\)](#) holds, consistency for the corresponding CATE would hold provided the other assumptions for consistency also hold for this restricted population.

Assumption 3. (*Limited Dependence for Consistency*) The joint distribution of treatment indicators $\{D_{n,i}\}_{i=1}^{N_n}$ satisfies

$$\frac{1}{N_n^2} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |\text{Cov}(D_{n,i}, D_{n,j})| = \frac{1}{N_n^2} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |\pi_{n,ij} - \pi_{n,i}\pi_{n,j}| = o(1) \quad (1.5)$$

Like [Assumption 2](#), this is also an implicit joint restriction on the ingredients of the spatial causal inference problem: the sequences of domains, units locations, and spatial assignment mechanisms. For example, under the treatment rule $D_i = \mathbf{1}\{d(\ell_i, \{s_c\}_{c=1}^C) \leq \delta\}$, this rules out the case where δ is so large (relative to the size of the domain) that it is possible for a single treatment center to treat all units. That would imply $\pi_{n,ij} - \pi_{n,i}\pi_{n,j} > 0$ for all pairs, implying the above quantity is not $o(1)$.

Intuitively, this assumption holds when the configuration of units within the domain plus the assignment mechanism together imply fuzzy clustering that is limited in scope. This occurs, roughly, in settings where the domains are growing asymptotically and contain increasingly many units, treated by increasingly many treatment centers, with units' treatment status independent when they are "far enough" apart. As an example sufficient condition, [Equation \(1.5\)](#) would be satisfied if any given unit's treatment is correlated with, at most, a finite number of other units. Then $\pi_{n,ij} - \pi_{n,i}\pi_{n,j} \neq 0$ for at most finitely many units so that the double sum in [Equation \(1.5\)](#) is $O(N_n^{-1}) = o(1)$.

Proposition 2. *Given [Assumption 1](#)-[Assumption 3](#),*

$$\begin{bmatrix} \hat{\mu}_{n,Y(1)} - \mu_{n,Y(1)} \\ \hat{\mu}_{n,Y(0)} - \mu_{n,Y(0)} \\ \hat{R}_{n,1} \\ \hat{R}_{n,0} \end{bmatrix} \xrightarrow{p} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

As a consequence $\hat{\alpha}_{n,ATE} - \alpha_{n,ATE} \xrightarrow{p} 0$.

Proof. See [Appendix A.3.3](#). □

1.3.3 Asymptotic Normality

This section proves asymptotic normality for the WLS estimator under the explicitly spatial central limit theorem of [Jenish and Prucha \(2009\)](#). This result shows explicitly how, within the bipartite spatial causal setting, the assumption of spatial weak dependence in assignment can be used to generate asymptotic normality that justifies inference. This is therefore an extension of [Aronow et al. \(2017\)](#), better adapted to spatial settings by making use of a spatial CLT.²¹ Additional details and comparisons of the conditions and available limit theorems are available in [Appendix A.2](#).

Assumption 4. (*Expanding Domain, No Infill Asymptotics*) There exists a finite constant $d_{min} > 0$ such that the distance between any pair of units $d(\ell_{n,i}, \ell_{n,j}) > d_{min}$ for all n and for all pairs of units i, j within the n th population.

By assuming a minimum distance we are ruling out infill asymptotics and implicitly restricting attention to sequences of experiments within large domains S_n accommodating increasingly many units spread across the domain. This rules out infinitely many units or observations piling up within any fixed-size neighborhood within the non-increasing domain. Expanding domain asymptotics are appropriate when studying discrete economic units. Infill asymptotics are most commonly used in mining and agricultural applications, where one is interested in learning about some continuous surface of outcomes; see [Diggle \(2013\)](#).

The next definition and assumption strengthen weak dependence [Assumption 3](#) in order deliver asymptotic normality.

²¹[Aronow et al. \(2017\)](#) use the CLT in [Chen and Shao \(2004\)](#), which is tougher to verify in this spatial context and less simple than distance-based decay, as in [Jenish and Prucha \(2009\)](#).

Definition 6. (Spatial Mixing Coefficient) For the the triangular sequence of random treatment indicators $\{D_{n,i}\}_{i=1}^{N_n}$, define the following mixing coefficients,

$$\alpha_{k,p}(m) = \sup_n \left[\sup_{|U| \leq k, |V| \leq p, d(U,V) \geq m} \left[\sup_{A \in \sigma(\{D_{n,i}\}_{\ell_i \in U}), B \in \sigma(\{D_{n,i}\}_{\ell_i \in V})} |P[A \cap B] - P[A]P[B]| \right] \right]$$

where $d(U, V) = \inf\{d(\ell, \ell') \mid \ell \in U, \ell' \in V\}$ (the smallest distance between two subsets of units) and $d(\ell, \ell')$ is Euclidean distance (see [Definition 33](#) for more on both).

Intuitively, $\alpha_{k,p}(m)$ is a numerical representation of the maximum amount of dependence between sets of k and p units that are at least distance m apart. Proceeding in pieces, recall that $P[A \cap B] - P[A]P[B]$ is a measure of dependence, as usual, equal to zero when events A and B are independent, nonzero otherwise. Therefore, the innermost expression measures the maximum amount of dependence between the set of random variables of $\{D_{n,i}\}$ at locations $U \subseteq \mathcal{I}$ and the random variables of $\{D_{n,i}\}$ at locations $V \subseteq \mathcal{I}$. The middle supremum ensures that the sets U and V contain no more than k and p elements and are at least distance m away under the metric $d(U, V)$. Finally, the outermost supremum looks at the maximum over all finite populations, which is necessary in the triangular setting. See [Definition 35](#) in [Appendix A.2](#) for more details.

Remark. While this paper focuses on spatial assignment and Euclidean distance as the metric in which dependence decays, one could also use other metrics or notions of distance. The exact choice is not important, but rather that correlation in the binary treatment indicators decays with some metric.

Assumption 5. (Weak Dependence Assumptions on Treatment) Suppose that the triangular array $\{D_{n,i}\}_{i=1}^{N_n}$ satisfies the following weak dependence conditions.

(i) For $k + \ell \leq 4$,

$$\sum_{m=1}^{\infty} m^{d-1} \cdot \alpha_{k,\ell}(m) < \infty$$

where $d = 2$. In words, the dependence between any two units, any unit and another pair, any pair, or any unit and three units must decay quickly enough with distance.

(ii) For some $\varepsilon > 0$,

$$\alpha_{1,\infty}(m) = O(m^{-2-\varepsilon})$$

In words, the dependence between any given unit and any arbitrarily large set of other units must

decay quickly enough with distance.

(iii) For some $\delta > 0$ and $d = 2$,

$$\sum_{m=1}^{\infty} m^{d-1} \alpha_{1,1}(m)^{\delta/(2+\delta)} < \infty$$

These conditions also imply $\frac{\delta}{2+\delta} < 1$ and $\alpha_{1,1}(m) \rightarrow 0$.

Because $\alpha_{k,p}(m)$ essentially measures dependence among (groups of) units at least distance m away, these restrictions require that dependence decays fast enough with distance so that all “sufficiently separated” units are not too correlated.

These conditions are slightly stronger than necessary for applying the CLT of [Jenish and Prucha \(2009\)](#) but are chosen to more closely match the weak dependence conditions familiar from [Conley \(1999\)](#) and [Bolthausen \(1982\)](#). The latter papers prove asymptotic normality in a random sampling context, where *outcomes* are draws from a *stationary* random field exhibiting weak dependence. Because I wish to accommodate arbitrary heterogeneity in potential outcomes, I use design-based uncertainty within a triangular asymptotic framework in which only the treatment indicators $\{D_{n,i}\}_{i=1}^{N_n}$ are random and exhibit weak dependence, while the ATE estimator is expressed as an average of rescaled random treatment indicators. Because of this difference, I need a central limit theorem that accommodates spatial weak dependence, non-stationarity of the terms in the sum defining the ATE, and triangular asymptotics. [Jenish and Prucha \(2009\)](#) offer one such CLT with conditions that are very related to the conditions in [Bolthausen \(1982\)](#), though applied to treatment as in [Abadie et al. \(2020\)](#).

Finally, weakening these conditions would further come at a small technical cost that is likely impractical for causal inference applications, unless a specific assignment process does not fit within this framework. However, see [Appendix A.2](#) for a more detailed discussion of mixing conditions and paths to weakening.

Proposition 3. *Conveniently,*

- (i) *Suppose that $d(\ell_i, \ell_j) > d_{\max}$ implies that $D_{n,i} \perp D_{n,j}$. Then [Assumption 5](#) is satisfied.*
- (ii) *[Assumption 5](#) and [Assumption 4](#) together imply [Assumption 3](#).*

Proof. See [Appendix A.3.4](#). □

Definition 7. (Randomization Variance) For each n , define

$$V_n := \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{e_{n,i}(1)}{\pi_{n,i}} + \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \right] \left[\frac{e_{n,j}(1)}{\pi_{n,j}} + \frac{e_{n,j}(0)}{1 - \pi_{n,j}} \right] [\pi_{n,ij} - \pi_{n,i}\pi_{n,j}]$$

$$e_{n,i}(d) = Y_{n,i}(d) - \mu_{n,Y(d)} \quad d \in \{0, 1\}$$

Notice that the variance depends upon *joint* treatment probabilities for pairs of units. Because of “fuzzy” clustering, the variance is inflated with additional terms.

Assumption 6. (*Nondegenerate Variance*) Assume there exists a fixed constant $b > 0$ such that, for all n ,

$$V_n > b$$

This rules out knife-edge assignment mechanisms and zero treatment effects.

I can now state the asymptotic normality result, which will form the basis of inference and motivate the upcoming comparison of variance estimators for the variance implied by the following proposition.

Proposition 4. *Under Assumptions 1–6,*

$$\sqrt{N_n} V_n^{-1/2} (\hat{\alpha}_{n,ATE} - \alpha_{n,ATE}) \xrightarrow{d} \mathcal{N}(0, 1)$$

Proof. See [Appendix A.3.4](#). □

Remark. This result nests a special case of [Abadie et al. \(2020\)](#), who similarly prove asymptotic normality conditional on potential outcomes, but in the case of individualistic assignment. The result of this section is suitable for handling the more general case of fuzzy clustering with spatial weak dependence.

1.4 Variance Estimation

This section states the suggested variance estimator and demonstrates that it is preferable to [Conley \(1999\)](#) standard errors and SHAC estimators more generally. I do this by deriving the limit of the alternative variance estimators under weak dependence on the treatment indicators, with only design-based uncertainty. A comparison of the resulting limits demonstrate that [Conley \(1999\)](#) standard errors are likely too large and lead to overcoverage (under positive correlation in potential outcomes, i.e. homophily) even in large samples.

In finite samples, [Conley \(1999\)](#) standard errors can also tend to exhibit undercoverage under unfavorable configurations of potential outcomes.

1.4.1 Estimators

In this section, I define the alternative variance estimators under consideration. The first makes use of [Aronow et al. \(2017\)](#), which exploits knowledge of individual *and* joint treatment probabilities. The second estimator is the SHAC class that includes [Conley \(1999\)](#) and makes use only of individual treatment probabilities along with a kernel weighting function.

Definition 8. (Modified Aronow and Samii Variance Estimator) Define the following variance estimator, which follows [Aronow et al. \(2017\)](#):

$$\begin{aligned}\hat{V}_{n,AS} = & \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{\pi_{n,i}\pi_{n,j}} - \frac{1\{i \neq j\}}{\pi_{n,ij}} \right] (Y_{n,i} - \hat{\mu}_{n,Y(1)})(Y_{n,j} - \hat{\mu}_{n,Y(1)})D_{n,i}D_{n,j} \\ & + \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{(1-\pi_{n,i})(1-\pi_{n,j})} - \frac{1\{i \neq j\}}{1-\pi_{n,i}-\pi_{n,j}+\pi_{n,ij}} \right] (Y_{n,i} - \hat{\mu}_{n,Y(0)})(Y_{n,j} - \hat{\mu}_{n,Y(0)})(1-D_{n,i})(1-D_{n,j}) \\ & + 2 \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j \neq i}^{N_n} \frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{\pi_{n,i}(1-\pi_{n,j})(\pi_{n,i} - \pi_{n,ij})} (Y_{n,i} - \hat{\mu}_{n,Y(1)})(Y_{n,j} - \hat{\mu}_{n,Y(0)})D_{n,i}(1-D_{n,j})\end{aligned}$$

While lengthy, the expression is written entirely in terms of observed outcomes, observed treatment indicators, and first- and second-order treatment probabilities.

This estimator is “modified” in the sense that I use $Y_{n,i} - \hat{\mu}_{n,Y(d)}$ in place of just $Y_{n,i}$. This is important because it’s what’s needed to mimic V_n in [Definition 7](#). Failing to subtract $\hat{\mu}_{n,Y(d)}$ and using the variance estimator of Expression (11) in [Aronow et al. \(2017\)](#) would lead to a non-straightforward comparison with [Conley \(1999\)](#), which automatically works with the regression errors (which are demeaned). This is one reason why [Leung \(2022\)](#) did not find a ranking between [Conley \(1999\)](#) and [Aronow et al. \(2017\)](#) in Appendix B, while I do.

Next, I consider subclasses of spatial heteroskedasticity and autocorrelation consistent (SHAC) variance estimators—specifically kernel-based SHAC estimators—versus the estimator of [Definition 8](#). This restriction to kernel-based SHAC estimators is essentially without loss of generality.²² This restriction also allows me to illustrate within the simpler kernel setting the problems associated with SHAC variance estimators more broadly, in comparison to the estimator of [Definition 8](#) that exploits the randomization distribution for

²²As in the time series literature, while there exist alternative SHAC estimators based on the periodogram in the frequency domain, there is an equivalence between kernel-based estimation and frequency-domain estimation of the variance, up to approximation error. See [Müller \(2014\)](#) and [Lazarus et al. \(2018\)](#).

design-based inference. The focus of this paper is on that difference, rather than in comparing the relative merits of alternative SHAC estimators within a causal setting.

Therefore, as a first step, I define this broad class of kernel variance estimators. Kernel variance estimators weight the cross product of WLS regression residuals by some kernel-based weights that are a nonnegative, weakly-declining function of the distance between units within the pair. Estimators within this class account for spatial correlation by including within the “meat” of a regression sandwich estimator additional terms for pairs of units that are spatially proximate (up to some cutoff controlled by a bandwidth choice) and, hence, likely correlated either due to assumed correlation in outcomes or correlation in treatment due to spatial assignment.

Definition 9. (Kernel Class of Variance Estimators) Let $K(u)$ denote a function satisfying

- (*Nonnegative*): $K(u) \geq 0$
- (*Symmetry*): $K(u) = K(-u)$
- $K(0) = 1$ and $K(u) = 0$ for $|u| > 1$, and bounded above by 1 for all u .

For some nonnegative bandwidth sequence b_n , let

$$K_{n,ij} = K(d(\ell_{n,i}, \ell_{n,j})/b_n)$$

with $d(\ell, \ell')$ measuring the Euclidean distance between any two locations ℓ and ℓ' .

Given the regression estimator $\hat{\beta}_{n,WLS}$ described in [Definition 4](#), define the corresponding WLS regression residuals $\hat{e}_{n,i} = Y_{n,i} - X'_{n,i}\hat{\beta}_{n,WLS}$. Then with these residuals, the kernel $K(u)$, and bandwidth sequence $b_n > 0$, define the corresponding kernel-based variance estimator for n th population as

$$\hat{V}_{n,K,b_n} = N_n \begin{pmatrix} -1 & 1 \end{pmatrix} (X'_n \omega_n X_n)^{-1} \left[\sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \omega_{n,i} \omega_{n,j} \hat{e}_{n,i} \hat{e}_{n,j} X_{n,i} X'_{n,j} K_{n,ij} \right] (X'_n \omega_n X_n)^{-1} \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

where $\omega_n = \text{diag}(\omega_{n,1} \cdots \omega_{n,N_n})$ using the weights from [Definition 4](#). Because the estimator depends upon choice of kernel function $K(u)$ and bandwidth b_n , this is a class of variance estimators indexed by the kernel K and b_n sequence.

Within the above class of kernel variance estimators, I consider a few special cases.

Definition 10. (Uniform Kernel, with Fixed-Bandwidth) This estimator corresponds to a uniform kernel $K(u) = \mathbf{1}\{|u| \leq 1\}$ with a fixed bandwidth $b_n = b$.

This variance estimator is most appropriate (although not ideal, as will be shown) for assignment mechanisms such that correlation of treatment indicators is zero beyond some fixed distance b , even asymptotically.

Definition 11. (Any Kernel, with Increasing-Bandwidth) The leading estimator proposed by [Conley \(1999\)](#) corresponds to a kernel satisfying [Definition 9](#) and bandwidth sequence satisfying $b_n = o(N_n^{1/6})$. This estimator asymptotically puts weight on all residual cross-products.

Remark. (Importance of Weighting) The original Conley estimator did not weight by treatment probabilities using ω_n . However, I do because, as stressed in [Borusyak and Hull \(2023\)](#), one cannot typically consistently estimate treatment effects without adjusting for differential treatment probabilities as is common in spatial settings.

Remark. (Bandwidth Choice) The choice of bandwidth $b_n = o(N_n^{1/6})$ was selected assuming random sampling of outcomes from a stationary field superpopulation (not a design-based perspective) and to ensure consistency under the weak dependence assumptions and proof strategies employed by Conley. There is, in general, no reason that this should be an appropriate way to construct a variance estimator if one is interested in design-based inference.²³

Remark. (Accounting for Correlation in Treatment) While both estimators from [Definition 8](#) and [Definition 9](#) use first order treatment probabilities $\pi_{n,i}$ to implement inverse-probability weighting, *only* the estimator of [Definition 8](#) makes use of the second order treatment probabilities $\pi_{n,ij}$ directly. Instead, the estimator of [Definition 9](#) uses the kernel as a blunt way of accounting for correlation, which drives its undesirable coverage properties. Under known assignment mechanism, it leaves on the table information (i.e. the $\pi_{n,ij}$ terms) that can sharpen inference.

Both kernel-based variance estimators above used the uniform kernel, and hence are not guaranteed to be positive in finite samples. The following restricted subclass of kernel estimators corrects that.

Definition 12. (PSD Class of Variance Estimators) In addition to the conditions in [Definition 9](#), assume

²³If one *does* wish to take a random sampling perspective, the [Conley \(1999\)](#) estimator does also seem to be suboptimal for inference. Specifically, the $o(N_n^{1/6})$ choice of bandwidth stemmed from using the weak dependence assumptions and covariance inequality from [Bolthausen \(1982\)](#) along with the proof strategy of [Quah \(1990\)](#). Under this combination, the bandwidth parameter can grow only at $o(N_n^{1/6})$. However, this rate for the two-dimensional setting corresponds to $o(N_n^{1/3})$ in the one-dimensional or time series setting, which is (1) slower than the $O(N_n^{1/3})$ MSE-optimal rate of [Andrews \(1991\)](#), (2) slower than the most recent guidance that recommends a larger bandwidth still in practice ([Lazarus et al., 2018](#)), and (3) slower than the slowest permitted rate $o(N_n)$ shown by [Andrews \(1991\)](#) for the time series case.

- (*Continuity*): $K(u)$ is continuous over its domain
- (*Nonnegative Spectral Generator Window*): The inverse Fourier transform of the kernel is everywhere nonnegative,

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} K(u) e^{-iu\lambda} du \geq 0$$

This definition restricts to (a subclass of) $\mathcal{K}_2$ in [Andrews \(1991\)](#). The two-dimensional spatial setting is analogous to the one-dimensional time series setting.

1.4.2 Comparison of Variance Estimators

This section states the main result, which makes clear that [Conley \(1999\)](#) and SHAC variance estimators more generally are inferior to the estimator $\hat{V}_{n,AS}$ that fully exploits the assignment mechanism. SHAC estimators are typically excessively conservative for design-based inference, but the exact degree of bias depends upon the inherently unobservable spatial correlation of heterogeneous treatment effects. There is no correction for SHAC estimators, except to use an explicitly design-based variance estimator.

Note that this ranking of estimators is a precise characterization under the particular assumptions I employ in this paper, which include no interference and restrictions on the mapping from spatial assignment to treatment. However, the insights generated also suggest extensions to the network case with interference, under correlated assignment, which I comment on further.

Assumption 7. (Strengthening of Weak Dependence Assumption) For some $\varepsilon > 0$,

$$\alpha_{\infty,\infty}(m)^{\delta/(2+\delta)} = O(m^{-(4+\varepsilon)})$$

This is a strengthening of [Assumption 5\(ii\)](#).²⁴ It was not necessary for normality, but is needed for consistency of [Definition 8](#), under my approach. It is not clear that it can be weakened.

Proposition 5. *Under Assumptions 1-7,*

$$\hat{V}_{n,AS} - V_n - \underbrace{\frac{1}{N_n} \sum_{i=1}^{N_n} (e_{n,i}(1) - e_{n,i}(0))^2}_{\text{Inestimable Component}} = o_p(1) \quad (1.6)$$

²⁴It is also the analog of C2 in [Conley \(1999\)](#).

where V_n is the variance from [Definition 7](#).

Proof. See [Appendix A.3.5](#). A similar result exists in [Aronow et al. \(2017\)](#), who proved the result under alternative assumptions distinct from weak spatial dependence. $\square$

Remark. Notice that $\widehat{V}_{n,AS}$ is consistent not for the true variance, but the true variance *plus* an inherently inestimable *positive* quantity equal to the the population variance of individual treatment effects.²⁵ For a given finite sample, conservative variance estimation is a general feature of causal inference, when using only randomization of assignment over the full sample rather than random sampling for inference. Even in the simplest textbook case of randomly assigning a fixed fraction of units to treatment, standard variance estimators are usually only *conservative* for the true variance, unless one appeals to random sampling from a larger superpopulation, which is undesirable in this spatial setting.²⁶ This is a manifestation of the usual missing data problem of causal inference for a given finite population. We cannot ever observe counterfactuals for the same unit, and conservativeness follows.

Moreover, this inestimable component will not vanish asymptotically, precisely because I adopt triangular asymptotics. Rather than sampling more units from a superpopulation, the asymptotics I employ are a sequence of experiments, each with the full population. Therefore, the inestimable component and inherent conservativeness of the variance estimator never disappear. See also [Aronow et al. \(2017\)](#) for discussion.

Finally, it is important to note that the conservativeness of $\widehat{V}_{n,AS}$ equals the conservativeness in the independent assignment case. In other words, even with correlated assignment, proper adjustment to the variance estimator ensures that the degree of conservativeness need not get worse.

Proposition 6. ($\widehat{V}_{n,C}$ under Increasing Bandwidth) *Suppose Assumptions 1-7 hold, that the kernel satisfies [Definition 9](#), and that $b_n = o(N_n^{1/6})$. This is the leading variance estimator of [Conley \(1999\)](#), reproduced in [Definition 11](#). Then*

$$\widehat{V}_{n,C} - V_n - \underbrace{\frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} [e_{n,i}(1) - e_{n,i}(0)][e_{n,j}(1) - e_{n,j}(0)] K_{n,ij}}_{\text{Inestimable Component} + \text{Additional Terms}} = o_p(1)$$

Proof. See [Appendix A.3.5](#). $\square$

²⁵More precisely, the variance of the dispersion of individual treatment effects around the overall mean since $e_{n,i}(1) - e_{n,i}(0) = (Y_{n,i}(1) - Y_{n,i}(0)) - (\mu_{n,Y(1)} - \mu_{n,Y(0)})$.

²⁶See [Imbens and Rubin \(2015\)](#) for this simple case. See also [Abadie et al. \(2020\)](#) for the generalization to regression with individualistic treatment assignment.

Corollary 1. ($\widehat{V}_{n,C}$ under Fixed Bandwidth) Suppose that one adopts a uniform kernel $K(u) = \mathbf{1}\{|u| \leq 1\}$ and fixed bandwidth $b_n = b$ so that

$$K_{n,ij} = \mathbf{1}\{d(\ell_i, \ell_j) \leq b\},$$

which places positive and uniform weight on any pair of units within distance b of each other. Then provided $d(\ell_i, \ell_j) > b$ implies $D_{n,i} \perp D_{n,j}$, it is the case that

$$\widehat{V}_{n,K_u,b} - V_n - \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j: d(\ell_i, \ell_j) \leq b} [e_{n,i}(1) - e_{n,i}(0)][e_{n,j}(1) - e_{n,j}(0)] = o(1)$$

Remark. (Use Case for Hard Truncation) If one knows that treatment is uncorrelated past a certain distance, then use of the uniform kernel and a fixed bandwidth does ensure, first, that one can conservatively estimate the true variance and, second, that the unobservable variance plus covariance portion does converge (the latter following in part from [Assumption 4](#), expanding domain asymptotics).

However, note that if the treatment indicators satisfy the more general weak dependence assumptions used in [Section 1.3.3](#), hard truncation may be inappropriate and lead to inconsistent estimation.

Remark. (Extra Terms in $\widehat{V}_{n,C}$) Similar to $\widehat{V}_{n,AS}$, the estimator $\widehat{V}_{n,C}$ is consistent for the truth V_n plus an inestimable quantity. However, relative to $\widehat{V}_{n,AS}$, this inestimable quantity includes additional terms, representing now the spatial *correlation* in heterogeneous treatment effects—not just their variance.²⁷ Therefore, comparison of the estimators amounts to a comparison of the additional terms.

Conveniently, the extra terms in each estimator do not even depend upon the spatial assignment mechanism, but rather the spatial distribution of individual treatment effects. Notice that the treatment probabilities π_i and π_{ij} are absent from the inestimable components. The inestimable terms come instead from the irresolvable missing data problem of causal inference and how each estimator handles it. As a consequence, only the spatial distribution of individual treatment effects and choice of kernel matter for the ranking of estimators.

As I now discuss, the inestimable portion can be positive or negative for $\widehat{V}_{n,C}$, depending on inherently unobservable spatial distribution of individual treatment effects.

Remark. (Usually too Conservative) Under positive spatial correlation in treatment effects, the double sum will be positive. This implies that the variance estimator is too large, confidence intervals are too long, and

²⁷ $e_{n,i}(1) - e_{n,i}(0) = (Y_{n,i}(1) - Y_{n,i}(0)) - (\mu_{n,Y(1)} - \mu_{n,Y(0)})$ is the demeaned treatment effect. Given that the kernel implements high weight on nearby units, this double sum captures spatial correlation.

one has overcoverage relative to $\widehat{V}_{n,AS}$. The stronger the spatial correlation in treatment effects, the more severe this problem becomes. See [Section 1.5](#).

Remark. (Possibly too Liberal) While it is often reasonable in applications to expect positive spatial correlation in treatment effects and thus conservativeness from $\widehat{V}_{n,C}$, it is still true that unobservable individual treatment effects could be chosen to render $\widehat{V}_{n,C}$ too liberal when using a fixed bandwidth. Said another way, for any n in the asymptotic sequence, one could find a particular configuration of potential outcomes and individual treatment effects such that the double sum in Proposition 1 is negative. As a result, $\widehat{V}_{n,C}$ would typically be smaller than V_n , even across repeated randomization over the assignment mechanism, for that adversarially chosen set of potential outcomes.

This result will be familiar from the time series literature and the variance estimator [Newey and West \(1986\)](#). That estimator was introduced to guard against unfavorable realizations from a time series that would imply a negative long run variance estimator, under the uniform kernel. The estimator $\widehat{V}_{n,C}$ uses the uniform kernel, and inference conditions on a realization of potential outcomes that could always be chosen unfavorably. In fact, in the triangular asymptotic framework (see [Section 1.2.4](#)), one could construct a sequence of unfavorable potential outcome functions for all n in the asymptotic sequence. And so even asymptotically, the problem does not diminish because one is averaging over randomization and not potential outcomes, which could be unfavorably chosen.

While this case is extreme and not always relevant in practice, it does highlight that the estimator $\widehat{V}_{n,C}$ is not optimized for the inference problem at hand and makes poor use of the assignment mechanism.

Remark. (Inefficient Use of the Assignment Mechanism) The extra terms stem from the “blunt” way in which $\widehat{V}_{n,C}$ accounts for correlation in treatment. Rather than exploit full knowledge of the assignment mechanism and the second order treatment probabilities $\pi_{n,ij}$, the estimator $\widehat{V}_{n,C}$ ignores these quantities and guards against arbitrary correlation in outcomes unrelated to assignment. While this may be appropriate in settings with interference, spatial assignment *alone* is not enough to warrant this choice, especially given the behavior highlighted in the remarks above.

On the other hand, the estimator $\widehat{V}_{n,AS}$ does exploit knowledge of second order treatment probabilities, and accordingly reduces the inestimable portion and conservativeness in the variance to the same magnitude as the independent assignment case, i.e. to the variance of individual treatment effects.

Remark. (PSD Kernels) One solution to the “too liberal” problem under fixed bandwidth would be to choose a kernel from [Definition 12](#), rather than the uniform kernel. Then even with a fixed bandwidth, the inestimable term would be positive. However, while solving the “too liberal” problem, this would still

typically lead to excessively conservative inference under positive spatial correlation in treatment effects, which is the most relevant case in practice.

Remark. (Simplifying Expressions for the Double Sum) I did not directly state a limit for the double sum. This limit depends on the superpopulation data generating process that produced potential outcomes as well as the spatial configuration of units. However, recall that inference is valid *conditional* on the data generating process and configuration of units in space—whatever they may be. Therefore, I do not impose restrictions on either, in order to derive the most general expression. The remarks above also highlight that one need not in order to derive useful results. That said, see [Section 1.5](#), where I make additional simplifying assumption to derive cleaner analytical expression and push the intuition further.

Remark. (Role of Heterogeneity) It is important to qualify that all differences between estimators are predicated upon *heterogeneity* in individual treatment effects. Under constant treatment effects, all variance estimators will coincide in large samples, and the variance estimators are *not* conservative. This again matches the textbook case of fully random assignment ([Imbens and Rubin, 2015](#)).

1.5 Results and Simulations for the Analytical Example

This section uses the analytical example introduced in [Section 1.2.5](#) to illustrate the results of [Sections 1.3](#) and [1.4](#) and perform a simulation study. This example also provides a clean characterization of a spatial causal inference problem, since major considerations for inference are characterized by a small number of parameters.²⁸

The simulation study first reinforces the conclusion in [Section 1.4.2](#) that Conley and SHAC standard errors more broadly can be overly conservative in large samples with strong persistence or nonstationarity in heterogeneous treatment effects. However, I also document non-trivial finite-sample undercoverage using the variance estimator of [Aronow et al. \(2017\)](#) and [Definition 8](#). I examine the causes, the severity, and potential avenues for improvement in future research.

1.5.1 Weak Dependence

In the simulation study, I will appeal to the asymptotic results in [Sections 1.3](#) and [1.4](#) to explain results.

This section verifies that the analytical example defined in [Section 1.2.5](#) satisfies the conditions outlined

²⁸The biggest simplification is in assuming equal spaced units, while typical examples in the spatial literature exhibit unequal spacing. Spacing of units is an important consideration in non-design-based spatial inference ([Müller and Watson \(2022\)](#)), where units' locations may be randomly sampled and that randomness must be correctly incorporated into inference. However, the design-based setting instead conditions on units' locations, with variation only over assignment.

above for asymptotic normality. It also gives one example of how one could verify sufficient conditions for normality given a parametric design for assignment.

Proposition 7. *The analytical example with treatment probabilities according to Proposition 1 satisfies Assumption 3.*

Proof. From the discussion in Section 1.2.5, $\pi_{ij} - \pi_i\pi_j = 0$ if $d(\ell_i, \ell_j) > 2\delta$. On the other hand, given the spacing of units in the analytical example, for a given unit at ℓ_i , there are $\lfloor \delta/\Delta \rfloor$ units on either side of unit i that are close enough to be correlated.

$$\frac{1}{N_n^2} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |\pi_{n,ij} - \pi_{n,i}\pi_{n,j}| = \frac{1}{N_n^2} \sum_{i=1}^{N_n} \sum_{m=-\lfloor \delta/\Delta \rfloor}^{\lfloor \delta/\Delta \rfloor} |\pi_{n,i,i+m} - \pi_{n,i}\pi_{n,i+m}|$$

And by symmetry and the uniform placement of units,

$$\frac{1}{N_n^2} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |\pi_{n,ij} - \pi_{n,i}\pi_{n,j}| = \frac{1}{N_n} \sum_{m=-\lfloor \delta/\Delta \rfloor}^{\lfloor \delta/\Delta \rfloor} |\pi_{n,0m} - \pi_{n,0}\pi_{n,m}|$$

Substituting in the treatment probabilities from Proposition 1,

$$\begin{aligned} \frac{1}{N_n} \sum_{m=-\lfloor \delta/\Delta \rfloor}^{\lfloor \delta/\Delta \rfloor} |\pi_{n,0m} - \pi_{n,0}\pi_{n,m}| &= \frac{1}{N_n} \sum_{m=-\lfloor \delta/\Delta \rfloor}^{\lfloor \delta/\Delta \rfloor} \left| 1 - 2e^{-2\delta\lambda} + e^{-\lambda[2\delta+|m\Delta|]} - (1 - e^{-2\delta\lambda})^2 \right| \\ &= \frac{1}{N_n} \left[(e^{-2\lambda\delta} - e^{-4\lambda\delta}) + 2 \sum_{m=1}^{\lfloor \delta/\Delta \rfloor} (e^{-\lambda[2\delta+m\Delta]} - e^{-4\lambda\delta}) \right] \end{aligned}$$

The sum on the last line is a sum of a finite number of elements. Therefore, the entire expression is $o(1)$ given the division through by N_n . $\square$

Proposition 8. *The analytical example with the assignment mechanism described in Section 1.2.5 satisfies Assumption 5.*

Proof. Again, from the discussion in Section 1.2.5, $\pi_{ij} - \pi_i\pi_j = 0$ and $D_i \perp D_j$ if $d(\ell_i, \ell_j) > 2\delta$. Therefore, by Proposition 3(i), taking $d_{max} = 2\delta$, Assumption 5 (weak dependence for normality) is satisfied. One can easily and analogously prove that Assumption 7 is satisfied as well. $\square$

1.5.2 Comparison of Variance Estimators

The results and conclusions of [Section 1.4.2](#) simplify further under the analytical example of [Section 1.2.5](#), and the inestimable “bias” in the variance expression can be easily characterized exactly as spatial autocovariances at particular lags, as I show in this section. I establish and highlight this result because it enables a connection to the time series HAC literature. This allows one to reason about the performance of SHAC estimators by appealing to known results on variance estimators from that literature.

Definition 13. (Spatial Autocovariance) For the example of [Section 1.2.5](#), define the spatial autocovariance at spatial lag k :

$$\gamma_{n,k} := \frac{1}{N_n} \sum_{i=1}^{N_n} [e_{n,i}(1) - e_{n,i}(0)] [e_{n,\text{mod}(i+k,N)}(1) - e_{n,\text{mod}(i+k,N)}(0)]$$

Because all units live on a circle, the first and the last units are proximate, hence the $\text{mod}(i+k, N)$.

Remark. (Role of the Data Generating Process) The spatial autocovariances $\gamma_{n,k}$ are finite-sample autocovariances, well-defined for *any* data generating process for outcomes and treatment effects. And as emphasized, all inference and results in this paper are fully conditioned on and valid regardless of the data generating process.

That said, as n increases, these finite-sample autocovariances will increasingly resemble the population autocovariances in the superpopulation data generating process (DGP), under regularity conditions. Therefore, the conservativeness one “expects” from $\hat{V}_{n,AS}$ and $\hat{V}_{n,C}$ is ultimately determined by the covariances in the superpopulation (DGP). I use this reasoning in the remarks following [Corollary 2](#).

The next corollary is a restatement of results that further clarifies the extent to which spatial autocorrelation in individual treatment effects ultimately determines conservativeness of [Conley \(1999\)](#) and SHAC estimators.

Corollary 2. (Restatement of Limits for Variance Estimators) *The preferred variance estimator satisfies*

$$\hat{V}_{n,AS} - V_n - \gamma_{n,0} = o_P(1)$$

The SHAC variance estimator under the uniform kernel with increasing bandwidth satisfies

$$\widehat{V}_{n,C,1} - V_n - \sum_{k=-c(b_n)}^{c(b_n)} \gamma_{n,k} = o_P(1)$$

where $c(b_n)$ is the number of units on one side of any given unit falling within distance b_n , where distance is measured by arc length.

The SHAC variance estimator under the uniform kernel with fixed bandwidth satisfies

$$\widehat{V}_{n,C,2} - V_n - \sum_{k=-c(b)}^{c(b)} \gamma_{n,k} = o_P(1)$$

Proof. By Proposition 8, Assumptions 1-7 are satisfied. Therefore, these results are a consequence of Proposition 5, Proposition 6, and Corollary 1. $\square$

Remark. (Conservativeness of $\widehat{V}_{n,C}$ depends on “Long Run Variance”) Leaning on this analytical example’s equivalence to the time-series case, the inestimable term of $\widehat{V}_{n,C,1}$

$$\sum_{k=-c(b_n)}^{c(b_n)} \gamma_{n,k}$$

is effectively the long-run variance for individual treatment effects. This implies useful intuition about the expected performance of the variance estimator of Conley (1999) and SHAC estimators in design-based settings more broadly.

First, in large samples, we can reason about the conservativeness of Conley (1999) by positing a data generating process for individual treatment effects and referring to its long run variance. Highly persistent processes have large long-run variances, i.e. many non-negligible autocovariance terms $\gamma_{n,k}$; therefore, Conley (1999) can be quite conservative. Alternatively, a data generating process with negative autocovariances can have a smaller long-run variance, reducing the (expected, large-sample) gap between V and $\widehat{V}_{n,C,1}$ or $\widehat{V}_{n,C,2}$. Moreover, the large-sample difference between the performance of $\widehat{V}_{n,AS}$ and $\widehat{V}_{n,C,1}$ or $\widehat{V}_{n,C,2}$ is governed by the difference between the unconditional variance and long run variance of a process.

Second, the above restatement is still a characterization of *asymptotic* conservativeness, but we can also use knowledge from the HAC literature to understand performance in finite samples. Specifically, the finite-sample performance of $\widehat{V}_{n,C,2}$ will be governed by the the finite-sample analog $\sum_{k=-c(b)}^{c(b)} \widehat{\gamma}_{n,k}$. As noted in the HAC literature, sample autocovariances might be such that for a process with negative autocovariances,

Parameter	Description	Value
Δ	Spacing of units	1
$ S_n $	Circumference	$N_n \cdot \Delta$
δ	Influence radius	$\in (1, 2, 3)$
λ	Assignment intensity	0.1

Table 1.1: Parameters used for simulations

this sum is negative. Therefore, [Conley \(1999\)](#) with the truncated kernel might satisfy $\widehat{V}_{n,C,2} < V_n$ in a given finite sample. Stated simply, one may observe finite-sample undercoverage when individual treatment effects follow a data generating process that would, in the HAC literature, raise the probability of a negative variance estimator under the truncated kernel. The simulation studies to come do show that such DGPs may also pose a problem for $\widehat{V}_{AS}$. So $\widehat{V}_{n,C,2}$ is not unique in this respect, but is still not ideal.

1.5.3 Simulation Study

This section contains the main simulation results on the performance of the variance estimator of [Definition 8](#) versus [Conley \(1999\)](#), the reference estimator from the SHAC literature. Throughout, I assume the setup of the analytical example in [Section 1.2.5](#).

Assignment and Treatment Probabilities

Following [Section 1.2.5](#), all units are evenly spaced along a circle, with a homogeneous Poisson process governing the (random, uniform) assignment of treatment centers over the circle.

As a consequence, treatment probabilities follow [Proposition 1](#) and depend upon parameters of the assignment mechanism, which are listed in [Table 1.1](#). In words, only the circle size $|S_n|$ scales with the number of units, while all other parameters remain fixed as a function of the sample size (but may vary to study coverage or power). The parameter δ controls the number of neighbors a given unit is correlated with, which (given integer values) is $2\delta - 1$ units on each side. This is the parameter that I vary to determine how finite-sample performance changes with the degree of correlation in treatment, as this parameter controls correlation in treatment, given fixed λ . Finally, the assignment intensity (and, therefore, first-order treatment probabilities for each unit) is fixed at $\lambda = 0.1$. I do not vary this parameter, and it was calibrated—conditional on the desired range of δ —to ensure that we essentially do not observe no-treated or all-treated realizations in the simulations.

Data-Generating Processes

Section 1.4.2 emphasized that the asymptotic performance of SHAC estimators depends crucially upon the spatial correlation in heterogeneous treatment effects, which is a feature of the data generating process (DGP) for potential outcomes. While typically unknown and application-dependent, I will write down a simple DGP family that helps illustrate the performance of the alternative variance estimators in small and large samples.

Definition 14. (Trend-Stationary DGP Family) I first assume that potential outcomes satisfy

$$\begin{aligned} Y_{n,i}(0) &\sim \mathcal{N}(10, 1), \\ Y_{n,i}(1) &= \alpha_{n,i} + Y_{n,i}(0). \end{aligned}$$

I will place alternative assumptions directly on the distribution of $\alpha_{n,i}$ which dictates asymptotic performance of the variance estimators, according to Section 1.4.2. The distribution of $Y_i(1)$ follows as a consequence.

Specifically, I assume that potential outcomes evolve around the circle like a trend-stationary process in the time-series literature (see, e.g. (Hamilton, 1994)):

$$\alpha_{n,i} = \tau_{n,i} + u_{n,i}$$

where $\tau_{n,i}$ is a deterministic trend, while $u_{n,i}$ is a stationary process for any fixed n . For the trend, I will consider two cases.

- Constant: $\tau_{n,i} = c$
- Linear: $\tau_{n,i} = (c - 2) + 4i/N_n$

The constant c will be set to $c = 2$ when studying size control of the alternative estimators. For power calculations, c will vary and therefore parameterize effect size.

For the stationary component u_t , I consider three cases:

- Null: $u_t = 0$.
- MA(1) with parameters θ, σ
- AR(1) with parameters ρ, σ

In the MA and AR designs, I will also consider local-to-unity values for θ and ρ , to study performance under stronger persistence. In the AR(1) case, I consider $\rho = 1$ as a fully nonstationary specification.

Finally, note that the designs may induce a cutoff effect at the origin of the circle, whereby potential outcomes are discretely jump, depending upon $\tau_{n,i}$ and $u_{n,i}$. Additional simulations have suggested that this does not drive the results presented below.

Variance Estimators

Variance in t -statistic	Definition and description
Feasible	
$\hat{V}_{EHW}$	Eicker-Huber-White estimator. Entails no adjustment for spatial correlation.
$\hat{V}_{AS}$	Definition 8, Aronow et al. (2017) estimator
$\hat{V}_{C,b}$	Definition 10, Conley (1999) under uniform kernel with fixed bandwidth $b = 2\delta$
$\hat{V}_{C,b_n}$	Definition 11, Conley (1999) with $b_n = N_n^{1/6.01}$
Infeasible	
V	Definition 7, True variance of the ATE estimator
$\hat{V}_{AS,e}$	Follow Definition 8 like $\hat{V}_{AS}$, but replace estimated error terms $\hat{e}_{n,i}(d) = Y_{n,i} - \hat{\mu}_{n,Y(d)}$ with infeasible $e_{n,i}(d) = Y_{n,i} - \mu_{n,Y(d)}$.

Table 1.2: Variances used in computing the t -statistic

The simulation study computes coverage and power of a t -statistic-based test, using the ATE estimator of Section 1.3 in the numerator, and alternative feasible and infeasible denominators for the variance. The denominator variances and corresponding notation are summarized in Table 1.2.

For the feasible variance estimators, we consider all alternatives studied in Section 1.4. This includes Conley (1999) under different choices of bandwidth. The first, $\hat{V}_{C,b}$, uses a fixed bandwidth $b = 2\delta$. In this analytical example, Proposition 1 implies that treatment is independent when $d(\ell_i, \ell_j) \geq 2\delta$, and this fixed-bandwidth estimator exploits that. This helps limit the “extra terms” asymptotically, which were established in Section 1.4.2. Therefore, its performance it will not “add terms” as N_n grows. In contrast, I also use $\hat{V}_{C,b_n}$, which adopts a valid asymptotically-growing bandwidth $b_n = N_n^{1/6.01} = o_p(N_n^{1/6})$, satisfying the conditions of Conley (1999). For the values of δ and sample sizes I consider, this rate can imply $b_n < b$ in some cases. The important thing is to note that $\hat{V}_{C,b_n}$ can become more conservative (under positive spatial correlation in treatment effects) as it continues to add terms. Finally, I also include use of Eicker-Huber-White standard

errors as a baseline in which the researcher does not adjust the variance estimator for spatial correlation in any way, to demonstrate when the spatial nature of the inference problem manifests strongly.

Lastly, to better interpret the results of the simulation study and understand drivers of undercoverage, I also study size and power under infeasible variances in the denominator of the t -statistic. The first, V , is the genuine true randomization variance of the ATE estimator, written in [Definition 7](#). The next two infeasible estimators reside in between infeasible V and feasible $\widehat{V}_{AS}$. Recall that V and $\widehat{V}_{AS}$ differ for three reasons. First, V contains an inestimable component related to the cross-product of unobserved potential outcomes for the same unit. Second, V and $\widehat{V}_{AS}$ differ because $\widehat{V}_{AS}$ must use estimates of the means $\widehat{\mu}_{n,Y(d)}$ in the variance computation, rather than true $\mu_{n,Y(d)}$, this potentially introduces bias. This may be possible to correct, similar to routine bias-adjust of variance estimators in the independent observation and clustering settings (see [Cameron and Miller \(2015\)](#) for more on the latter). For this reason, I define $\widehat{V}_{AS,e}$ as an infeasible estimator between V and $\widehat{V}_{AS}$. This infeasible estimator $\widehat{V}_{AS,e}$ is constructed exactly as in [Definition 8](#), but replacing estimated error terms $\widehat{e}_{n,i}(d) = Y_{n,i} - \widehat{\mu}_{n,Y(d)}$ with the truth $e_{n,i}(d) = Y_{n,i} - \mu_{n,Y(d)}$. Third and lastly, $\widehat{V}_{AS}$ only uses a portion of the population in estimating the variance and covariance terms since not all units are observed simultaneously in treated and untreated states. While this is inescapable, I also would like to gauge the extent to which estimating the correlation specifically (using only a subset of the population) poses a problem. This correlation term is what makes this design-based setting unique relative to [Abadie et al. \(2020\)](#). Therefore,

Simulation Results

This section reports coverage and power when using a confidence interval based on inverting a t -statistic based test, under alternative feasible and infeasible variance estimators. This gives a numerical accounting of the way in which SHAC estimators are unsuitable in design-based inference. Importantly, I can also explore small-sample performance, since the results of [Section 1.4.2](#) are asymptotic. And as will be shown and discussed, finite-sample considerations do play an important role in the performance of variance estimators, including $\widehat{V}_{AS}$. These simulation results therefore also inform recommendations for future research

No Heterogeneity Baseline [Table 1.3](#) offers a first, baseline look at the coverage comparisons in this section. This table shows the coverage of confidence intervals for the finite-sample ATE α_n . I repeatedly draw potential outcomes and the treatment effect from the DGP of [Definition 14](#), under the assumptions that $\tau_{n,i} = c$ is constant for all units and $u_{n,i} = 0$ for all units. Given a draw of potential outcomes, I randomly

$\tau_{n,i}$ Constant $u_{n,i} = 0$	$\delta = 1$			$\delta = 2$			$\delta = 3$		
Var. in t -stat	N=250	500	1000	250	500	1000	250	500	1000
Feasible									
$\hat{V}_{EHW}$	0.940	0.942	0.949	0.947	0.946	0.953	0.947	0.949	0.952
$\hat{V}_{AS}$	0.926	0.936	0.947	0.934	0.941	0.950	0.932	0.943	0.949
$\hat{V}_{C,b}$	0.933	0.940	0.947	0.931	0.942	0.948	0.926	0.941	0.947
$\hat{V}_{C,b_n}$	0.928	0.938	0.945	0.939	0.942	0.948	0.934	0.947	0.948
Infeasible									
V	0.944	0.946	0.950	0.944	0.944	0.953	0.941	0.948	0.951
$\hat{V}_{AS,e}$	0.937	0.943	0.950	0.939	0.944	0.953	0.936	0.944	0.949

Table 1.3: **Coverage under Constant Treatment Effect Baseline** Coverage under DGP in Definition 14, with $\tau_{n,i}$ constant, with $c = 2$, $u_{n,i} = 0$. I vary δ , which controls the number of neighbors each unit's treatment status is correlated with (on either side) as well as sample size N . Coverage is computed across 5000 simulations.

$\tau_{n,i}$ Deterministic $u_{n,i} = 0$	$\delta = 1$			$\delta = 2$			$\delta = 3$		
Var. in t -stat	N=250	500	1000	250	500	1000	250	500	1000
Feasible									
$\hat{V}_{EHW}$	0.908	0.911	0.911	0.864	0.860	0.869	0.852	0.851	0.858
$\hat{V}_{AS}$	0.930	0.947	0.951	0.934	0.945	0.957	0.939	0.945	0.956
$\hat{V}_{C,b}$	0.956	0.962	0.965	0.979	0.981	0.986	0.988	0.995	0.993
$\hat{V}_{C,b_n}$	0.963	0.973	0.980	0.971	0.971	0.986	0.969	0.969	0.985
Infeasible									
V	0.948	0.951	0.949	0.953	0.945	0.953	0.947	0.946	0.952
$\hat{V}_{AS,e}$	0.946	0.955	0.954	0.948	0.952	0.960	0.948	0.950	0.959

Table 1.4: **Deterministic Heterogeneous Treatment Effects** Coverage under DGP of Definition 14, with $\tau_{n,i}$ Linear, with $c = 2$, $u_{n,i} = 0$. Coverage is computed across 5000 simulations.

assign treatment centers to locations on the circle, which generates realization for treatment status $D_{n,i}$ of each unit. From there, I construct the estimator $\hat{\alpha}_n$ and construct confidence intervals for α_n using the estimators (feasible and infeasible) described above by inverting the t -statistic based test using the normal reference distribution, with target nominal coverage of 95%. Table 1.3 reports coverage of these confidence intervals.

Table 1.3 emphasizes that the spatial problem only requires adjustment insofar as treatment effects exhibit heterogeneity, a point emphasized earlier in this paper and the literature. Notice that under constant treatment effects, even Eicker-White standard errors achieve target coverage, regardless of the degree of spatial correlation in treatment, which is controlled by δ . In contrast, the spatial variance estimators $\hat{V}_{AS}$, $\hat{V}_{C,b}$, and $\hat{V}_{C,b_n}$ display worse coverage in finite samples. This table also previews some undercoverage in smaller samples when using the feasible spatial estimators $\hat{V}_{AS}$, $\hat{V}_{C,b}$, and $\hat{V}_{C,b_n}$, which we will see again in other simulations. In summary, it is precisely the existence of treatment effect heterogeneity that drives the need for adjustment.

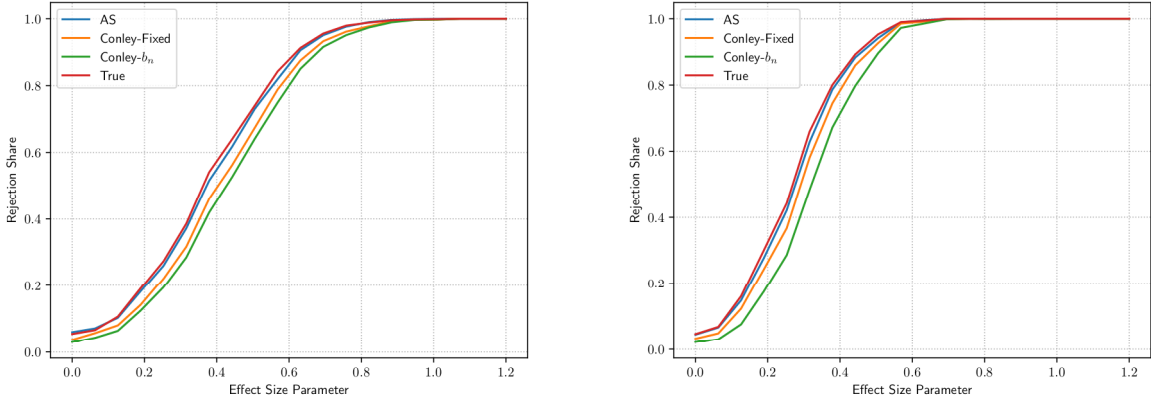


Figure 1.1: Power curves corresponding to the DGP of Table 1.4 and Definition 14, with $\tau_{n,i}$ Linear, with effect size c , $u_{n,i} = 0$. The effect size c is varied across values on the x -axis to trace out the rejection probabilities shown above for both $N = 500$ (left) and $N = 1000$.

Deterministic Treatment Effect Heterogeneity in Large and Small Samples Table 1.4 shows how treatment effect heterogeneity—even the simplest deterministic kind—manifests the results and conclusions of previous sections. I assume that $\tau_{n,i}$ varies linearly, inducing positive correlation in treatment effects. This is likely the most common and expected case; therefore, the results here are indicative of likely most applications in spatial causal inference.

First, Eicker-Huber-White standard errors deteriorate in coverage with greater correlation in treatment δ , as expected, which is evident by direct comparison with Table 1.3. Second, standard errors following Conley (1999), $\hat{V}_{C,b}$ and $\hat{V}_{C,b_n}$, start to exhibit conservativeness, as they are too large for design-based inference. To further emphasize the unsuitability of SHAC estimators for design-based inference, Figure 1.1 depicts power curves that demonstrate clear loss in power relative to the variance estimator $\hat{V}_{AS}$ of Aronow et al. (2017) and Definition 8. In a sample of $N = 1000$, the power loss of Conley (1999) estimator $\hat{V}_{b_n}$ using a sample-dependent bandwidth (green) is over 0.15 relative to $\hat{V}_{AS}$ (blue) in magnitude halfway towards rejection with probability one. Use of the truncated kernel in $\hat{V}_b$ lies in between. However, both are dominated in large samples by $\hat{V}_{AS}$. And the power loss of $\hat{V}_{AS}$ relative to true V (red) is quite minimal.

Lastly, Table 1.4 also suggests that $\hat{V}_{AS}$ controls coverage well despite the overall level of persistence, but appears to suffer from small-sample bias. Therefore, even though it dominates in larger samples by avoiding unnecessary conservativeness, this conservativeness may inadvertently help $\hat{V}_{C,b}$ and $\hat{V}_{C,b_n}$ in finite samples. Moreover, across δ , which controls the degree of correlation in treatment, $\hat{V}_{AS}$ performs similarly. Only the sample size appears to be of concern. I use the infeasible estimator $\hat{V}_{AS,e}$.

Stochastic Heterogeneity in Large and Small Samples Next, Tables 1.5 sets the deterministic portion of the DGP for treatment effects to a constant, $\tau_{n,i} = c$, but specifies an AR process for the stochastic component $u_{n,i}$. Results for the AR(1) case in Table 1.5 shares many of the same conclusions as for the deterministic $\tau_{n,i}$ case in Table 1.4. First, $\hat{V}_{C,b}$ and $\hat{V}_{C,b_n}$ are conservative. This is especially true for the most persistent DGPs with $\rho = 1 - 1/N$ or $\rho = 1$. Moreover, as N grows, This is consistent with the characterization of the conservative portion in Corollary 2, where the spatial autocovariances do not sharply decay and an increasing bandwidth simply adds unnecessary positive autocovariance terms on top of unnecessary positive autocovariance terms. This is an important takeaway for practice, as nonstationarity in treatment effects and potential outcomes is likely quite relevant in spatial settings (see Bera and Simlai (2005) and Banerjee et al. (2004)). The costs of the additional terms in SHAC estimators (in a design-based context) can be high in large samples. Moreover, the conservativeness of $\hat{V}_{C,b_n}$ will only grow with b_n and N_n .

In contrast, the MA(1) example with negative correlation highlights problems with $\hat{V}_{AS}$ and difficulties maintaining nonnegativity. While asymptotically conservative for V , $\hat{V}_{AS}$ need not be conservative or non-negative in finite samples. In the time-series HAC literature, nonnegativity of long run variance estimators is enforced by restricting the class of kernels (see Definition 12). However, $\hat{V}_{AS}$ is different as it is ultimately a Horvitz-Thompson type estimator of the variance. There is no guarantee or adjustments made to enforce nonnegative of this object. It is only unbiased for V plus another term to ensure asymptotic conservativeness. In a given sample, it can be negative. This is observed in 0% of simulations for $\theta = -0.5$, but approximately 12% for $\theta = -0.9$, and 20% for $\theta = -(1 - 1/N)$. Coverage values in the table for $\hat{V}_{AS}$ are conditional on positive values and therefore also minimize the issue.

$\tau_{n,i}$ Constant $u_{n,i} \sim AR(1)$	$(\rho, \sigma) = (0.9, 0.5)$			$(\rho, \sigma) = (0.99, 0.5)$			$(\rho, \sigma) = (1 - \frac{1}{N}, 0.5)$			$(\rho, \sigma) = (1, 0.5)$		
Var. in t -stat	N=250	500	1000	250	500	1000	250	500	1000	250	500	1000
Feasible												
$\hat{V}_{EHW}$	0.901	0.923	0.919	0.875	0.886	0.886	0.870	0.885	0.890	0.866	0.880	0.887
$\hat{V}_{AS}$	0.925	0.946	0.952	0.920	0.945	0.947	0.919	0.950	0.953	0.921	0.944	0.955
$\hat{V}_{C,b}$	0.949	0.964	0.966	0.948	0.969	0.969	0.950	0.971	0.974	0.954	0.971	0.975
$\hat{V}_{C,b_n}$	0.954	0.970	0.981	0.958	0.977	0.982	0.958	0.978	0.989	0.961	0.976	0.990
Infeasible												
V	0.944	0.952	0.950	0.937	0.949	0.945	0.936	0.952	0.948	0.939	0.950	0.950
$\hat{V}_{AS,e}$	0.939	0.954	0.956	0.935	0.954	0.951	0.935	0.957	0.957	0.937	0.951	0.959

Table 1.5: **Stochastic Heterogeneous Treatment Effects, AR(1)** Coverage under DGP in Definition 14, with $\tau_{n,i} = c$ with $c = 2$, $u_{n,i} \sim AR(1)$ under different persistence. Coverage is computed across 5000 simulations with $\delta = 1$.

$\tau_{n,i}$ Constant $u_{n,i} \sim MA(1)$	$(\theta, \sigma) = (-0.5, 0.5)$			$(\theta, \sigma) = (-0.9, 0.5)$			$(\theta, \sigma) = (-(1 - 1/N), 0.5)$		
Var. in t -stat	N=250	500	1000	250	500	1000	250	500	1000
Feasible									
$\hat{V}_{EHW}$	0.962	0.964	0.964	0.962	0.965	0.965	0.962	0.965	0.965
$\hat{V}_{AS}$	0.932	0.945	0.950	0.934	0.946	0.949	0.936	0.946	0.950
$\hat{V}_{C,b}$	0.928	0.941	0.945	0.928	0.942	0.945	0.928	0.943	0.945
$\hat{V}_{C,b_n}$	0.926	0.941	0.946	0.926	0.941	0.945	0.926	0.941	0.945
Infeasible									
V	0.943	0.949	0.948	0.943	0.948	0.947	0.944	0.948	0.947
$\hat{V}_{AS,e}$	0.944	0.951	0.951	0.945	0.951	0.952	0.946	0.951	0.952

Table 1.6: **Stochastic Heterogeneous Treatment Effects, MA(1)** Coverage under DGP in Definition 14, with $\tau_{n,i} = c$ with $c = 2$, $u_{n,i} \sim MA(1)$ under different MA coefficient. Coverage is computed across 5000 simulations with $\delta = 1$.

Conclusions and Discussion Broadly, the variance estimator of Aronow et al. (2017) appears to suitably handle correlation in treatment, but suffers from problems of small-sample bias that might be possible to remedy. While the estimator of Aronow et al. (2017) leads to approximate (only mildly conservative) nominal coverage in large samples (seemingly regardless of the degree in treatment correlation δ), positive spatial correlation in heterogeneous treatment effects appears to induce downward bias in estimates of the design-based variance. Conley (1999) and SHAC standard errors may suffer less in finite samples, but for the wrong reasons. They do not directly address the source of the undercoverage in design-based settings. They include extra additional terms, which are unnecessary from the perspective of design-based inference, lead to certain overcoverage in large samples (under positive correlation in treatment assignment), and possible undercoverage in finite samples. There is still a gap in the literature that could be filled with a proper small-sample bias-adjustment for the variance estimator of Aronow et al. (2017).

When there is negative correlation in heterogeneous effects, $\hat{V}_{AS}$ performs poorly. While this may not be the most common case in practice, it still could be remedied by adjustments to ensure $\hat{V}_{AS}$ is positive with probability one, as was done in the HAC literature (e.g. Newey and West (1986) and Andrews (1991)). The task would be to do so in a way that maintains a design-perspective and does not introduce unnecessary conservativeness, if possible.

As a way of decomposing and diagnosing the size loss, I included in all tables coverage under $\hat{V}_{AS,e}$. I included $\hat{V}_{AS,e}$ because, in the independent and identically distributed inference case, estimation of the variance using the sample mean causes known bias, motivating the common $N/(N - 1)$ correction. While such a correction is not immediately appropriate here with correlated treatment, $\hat{V}_{AS,e}$ uses the true mean directly, as a way of shutting down any bias contribution to $\hat{V}_{AS}$ from the use of estimated means $\hat{\mu}_{n,Y(d)}$ as opposed to true means. For $N = 250$, we see that this helps somewhat improve upon coverage using $\hat{V}_{AS}$ in

Table 1.5 (the stochastic AR(1) design), but not as much for deterministic treatment in Table 1.4.

Therefore, this simulation study recommends that variance estimation in design-based settings with correlated treatment might be a fruitful area of future research. Ideally, one could adjust for the poor finite-sample performance of explicitly design-based variance estimators like Definition 8, without sacrificing power in large samples, as SHAC estimators would imply in common settings.

1.6 Instrumental Variables under Joint Assignment with Treatment Effect Heterogeneity

In spatial or bipartite causal inference settings, instrumental variables are often employed to estimate causal effects, and I extend the results of the previous section to cover this case.²⁹ To my knowledge, this paper is the first to describe asymptotically normal inference for the local average treatment effect (LATE) under joint assignment that exhibits fuzzy clustering—i.e. the kind that arises in bipartite causal inference settings like the spatial case.³⁰ Recently, Borusyak and Hull (2023) also considered the IV setting and described inference for *constant* treatment effects, but they did not cover variance estimation under *heterogeneous* individual treatment effects. And while those others covered multivalued treatment, I limit this section to the binary case, although one could easily extend the results.

As in the ATE case, I will use a spatial CLT to justify inference. However, the same effect and variance estimators that I present could be used for inference in other fuzzy clustering settings and justified by alternative CLTs that best fit a given application. The particular LATE estimator and associated variance estimator are extensions of Aronow et al. (2017). The variance estimator will again exploit knowledge of the assignment mechanism, and the design-based inference perspective and results on conservativeness will carry hold in this setting as well.

1.6.1 Framework

This section starts from Section 1.2.1 and adds the additional details needed to cover IV estimation.

Spatial Assignment Mechanism Assume again a point process that describes the random allocation of treatment centers to points $\{s_c\}_{c=1}^C \subset S$. While this will not directly determine treatment (as in the previous

²⁹See Baum-Snow and Ferreira (2015) and Borusyak and Hull (2023) for examples where IV is employed in spatial settings and settings with joint assignment to treatment.

³⁰“Asymptotically normal inference” as opposed to inference based on Fisher randomization and sharp null hypotheses.

setting), the assignment mechanism will determine the realization of an instrument.

Potential Outcomes and Endogenous Regressor Not only does the IV setting have potential outcome functions $Y_i(\cdot)$, it also has an endogenous regressor, which I denote by $W_i(\cdot)$ and assume to be binary. Again, with treatment and outcomes occurring at different levels in the bipartite spatial setting, potential outcomes and endogenous regressors for unit i might depend upon the full realization of spatial treatment.

$$Y_i(\{s_c\}_{c=1}^C) \quad W_i(\{s_c\}_{c=1}^C)$$

However, analogous to the ATE setting, I will assume that the influence of the spatial assignment mechanism for the i th unit can be summarized by an individualistic binary instrument, which I denote

$$D_i = D_i(\ell_i, \{s_c\}_{c=1}^C)$$

so that the potential outcomes and endogenous regressor can be reparameterized as

$$Y_i(\{s_c\}_{c=1}^C) = Y_i(D_i) \quad W_i(\{s_c\}_{c=1}^C) = W_i(D_i)$$

Notice that D_i no longer plays the role of actual treatment, but rather an instrument or (in some settings) assignment to treatment, while W_i is the endogenous regressor.

While seemingly restrictive, this setup can accommodate reasonable and relevant experimental designs for spatial applications. For example, consider a setting where the spatial assignment mechanism locates resource centers at particular spatial locations $\{s_c\}_{c=1}^C$. Individuals are *able* to use the allocated resource if they are within some distance, which implies a well-defined D_i rule for each unit, depending upon their location and proximity to treatment. But not every unit ultimately takes up use of the resource if proximate to treatment, and so there is also an endogenous $W_i(D_i)$ for each unit, which could, for example, represent the notion of ultimate “take up” of treatment.

Treatment Probabilities As before, the spatial assignment mechanism implies first- and second-order treatment probabilities for the binary indicator D_i , which now represents the instrument as opposed to treatment:

$$\pi_i = P[D_i = 1] \quad \pi_{ij} = P[D_i = 1, D_j = 1].$$

Unlike standard IV settings, often $\pi_{ij} \neq \pi_i \pi_j$ because of joint assignment and the resulting fuzzy clustering.

1.6.2 LATE Estimand

This section will first define the estimand as a finite-sample Wald estimator, but then I state additional assumptions that give this estimand the interpretation of a finite-sample local average treatment effect (LATE). The necessary assumptions are the usual assumptions in an IV context—monotonicity, exclusion, relevance as in [Imbens and Angrist \(1994\)](#)—suitably adapted to this setting.

Importantly, the additional assumptions needed for a LATE interpretation do not restrict the assignment mechanism in any way or impose strong modeling assumptions on potential outcomes. These are substantive economic assumptions placed on the problem by the researcher to ensure identification of the estimand with a meaningful causal effect.

Definition 15. (Wald Ratio) Under the assumptions placed on potential outcomes and the endogenous regressor, one can define (for any finite sample) the Wald ratio

$$\alpha_{IV} := \frac{\alpha_{ATE,Y}}{\alpha_{ATE,W}} = \frac{\frac{1}{N} \sum_{i=1}^N Y_i(1) - \frac{1}{N} \sum_{i=1}^N Y_i(0)}{\frac{1}{N} \sum_{i=1}^N W_i(1) - \frac{1}{N} \sum_{i=1}^N W_i(0)}.$$

Initially, this does not have any particular meaning; however, the remainder of this section will place assumptions on $Y_i(d)$ and $W_i(d)$ that give this quantity meaning as a LATE.

Assumption 8. (Monotonicity) $W_i(1) \geq W_i(0)$ (or vice versa) for all units i in any population n in the sequence of populations. Under this assumption all units in the n th population can be divided into the following three standard groups:

- (i) Never-takers: $S_{NT} = \{i : W_i(1) = W_i(0) = 0\}$
- (ii) Always-takers: $S_{AT} = \{i : W_i(1) = W_i(0) = 1\}$
- (iii) Compliers: $S_C = \{i : W_i(1) > W_i(0)\}$

The monotonicity assumption rules out a potential fourth group of “defiers.”

Assumption 9. (Exclusion) D_i affects Y_i only through its effect on W_i , for any unit i in any population n in the sequence, i.e.

$$W_i(1) - W_i(0) = 0 \quad \implies \quad Y_i(1) - Y_i(0) = 0$$

This assumption ensures that assignment to treatment or control does not affect other factors than W_i that might impact potential outcomes.

Assumption 10. (Relevance, Existence of Compliers) There exist sufficiently many compliers,

$$\frac{1}{N_n} \sum_{i=1}^{N_n} [W_i(1) - W_i(0)] > k \quad \forall n$$

for some fixed $k > 0$. This is equivalently a relevance condition for the particular instrument D_i , ensuring that assignment to treatment has the ability to change the behavior of some units and that there is always a positive fraction of units that are compliers within the population.

Proposition 9. (LATE Estimands) *Under Assumptions 8-10, the Wald ratio α_{IV} of Definition 15 equals the average treatment effect among compliers,*

$$\alpha_{IV} = \alpha_{LATE} = \frac{1}{|S_{n,C}|} \sum_{i \in S_{n,C}} [Y_{n,i}(1) - Y_{n,i}(0)].$$

Proof. See Appendix A.4 for the proof, which mirrors Imbens and Angrist (1994). $\square$

1.6.3 Asymptotics and Inference for LATE Estimand

This section extends the results on consistency, asymptotic normality, and variance estimation from the ATE case in Sections 1.3 and 1.4 to the LATE case. Under the same weak dependence assumptions on D_i —now interpreted as an instrument rather than assignment—one can prove a CLT theorem and conservatively estimate the variance for the Wald estimator of Definition 15, under arbitrarily heterogeneous individual treatment effects and design-based uncertainty.

Proposition 10. *Define the following weights*

$$\omega_{n,i} = \frac{D_{n,i}}{\pi_{n,i}} + \frac{1 - D_{n,i}}{1 - \pi_{n,i}}$$

constructed from observed treatment status and individual treatment probabilities. Then let $\widehat{\beta}_{n,WLS,Y}$ and $\widehat{\beta}_{n,WLS,W}$ denote the estimators that result from two separate weighted least squares regressions of observed $Y_{n,i}$ on $X_{n,i} = (1 - D_{n,i}, D_{n,i})$ and $W_{n,i}$ on the same $X_{n,i}$, both using weights $\{\omega_{n,i}\}_{i=1}^{N_n}$. Also let

$$\widehat{\alpha}_{n,ATE,Y} := \begin{pmatrix} -1 & 1 \end{pmatrix} \widehat{\beta}_{n,WLS,Y} \quad \widehat{\alpha}_{n,ATE,W} := \begin{pmatrix} -1 & 1 \end{pmatrix} \widehat{\beta}_{n,WLS,W}.$$

Then define the following weighted Wald ratio estimator

$$\hat{\alpha}_{n,IV} = \frac{\hat{\alpha}_{n,ATE,Y}}{\hat{\alpha}_{n,ATE,W}}$$

Given Assumptions 1-3 and Assumptions 8-10,

$$\hat{\alpha}_{n,IV} - \alpha_{n,LATE} \xrightarrow{p} 0.$$

Proof. From Proposition 2, $\hat{\alpha}_{n,ATE,Y} - \alpha_{n,ATE,Y} = o_p(1)$. This proposition can also be applied to the endogenous variables to yield, $\hat{\alpha}_{n,ATE,W} - \alpha_{n,ATE,W} = o_p(1)$. Together with the continuous mapping theorem, the desired result follows. $\square$

Remark. This result relates to a contemporaneous result in the appendix of [Borusyak and Hull \(2023\)](#), who provide an alternative regression-based estimation approach for causal effects and show that their method, under treatment effect heterogeneity, also recovers the LATE $\alpha_{n,LATE}$. In particular, the authors' approach uses *unweighted* two-stage least squares regression of observed outcome Y_i on a constant, the regressor W_i instrumented by D_i , and an additional regressor equal to the treatment probability π_i .³¹ Regression matrix algebra shows that the coefficient on D_i in their approach is identical to the Wald ratio estimator $\hat{\alpha}_{n,LATE}$ presented above, which uses *weighted* least squares regression of Y_i on a constant and D_i , with weights given by the inverse of π_i . Both approaches yield the same estimator, consistently estimate the same quantity, and appropriately account for non-identical first-order treatment probabilities $\{\pi_i\}_{i=1}^N$, which is the main concern of [Borusyak and Hull \(2023\)](#). However, I will now derive the appropriate asymptotic variance under treatment effect heterogeneity and provide a variance estimator, a task which [Borusyak and Hull \(2023\)](#) left open, given that their main concern was regression-based estimation of *constant* causal effects.

Definition 16. (Randomization Variance for LATE) For each n , define

$$V_n := \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{e_{n,i}(1)}{\pi_{n,i}} + \frac{e_{n,j}(0)}{1 - \pi_{n,i}} \right] \left[\frac{e_{n,j}(1)}{\pi_{n,j}} + \frac{e_{n,j}(0)}{1 - \pi_{n,j}} \right] [\pi_{n,ij} - \pi_{n,i}\pi_{n,j}]$$

$$e_{n,i}(d) = \frac{Y_{n,i}(d) - \alpha_{n,IV} W_{n,i}(d) - [\mu_{n,Y(d)} - \alpha_{n,IV} \mu_{n,W(d)}]}{\mu_{n,W(1)} - \mu_{n,W(0)}} \quad d \in \{0, 1\}$$

³¹In the binary setting, $\pi_i = P[D_i = 1] = \mathbb{E}[D_i]$ equals precisely their “expected instrument,” which [Borusyak and Hull \(2023\)](#) show is necessary to include as an additional regressor in order to avoid bias whenever first-order treatment probabilities $\{\pi_i\}_{i=1}^N$ are heterogeneous and one uses unweighted two-stage least squares to estimate the causal effect. I avoid bias by instead using inverse probability weighting, which is numerically identical in this binary treatment case.

This is the analogue of [Definition 7](#) for the LATE case.

Proposition 11. *Assume that regularity and weak dependence Assumptions 1–6 hold.³² Suppose also that the substantive LATE Assumptions 8–10 hold. Then*

$$\sqrt{N}V_n^{-1/2}(\hat{\alpha}_{n,IV} - \alpha_{n,LATE}) \xrightarrow{d} \mathcal{N}(0, 1)$$

Proof. The proof is a direct extension of the proof of [Proposition 4](#). □

Remark. Notice that no further assumptions have been placed on the assignment mechanism, only on potential outcomes and the endogenous regressors in the form of [Assumption 8](#)–[Assumption 10](#).

The variance has nearly the exact same form as the variance in [Proposition 4](#) in the previous section, which immediately suggests a variance estimator. Specifically, the variance expression in this IV setting simply uses a different expression than [Definition 7](#) for $e_{n,i}(d)$. As a result, the same considerations and the same estimation approach apply. Namely, there will be portions of this variance that are not consistently estimable, as they involve within-unit products of counterfactual outcomes (which are never observable). However, because this variance has the same structure as the variance in [Proposition 4](#), I can solve the variance estimation problem in the same way, using the conservative variance estimator from [Aronow et al. \(2017\)](#). And because no further assumptions on the assignment mechanism have been made in the IV context, consistent variance estimation obtains under the same set of assumptions as in [Proposition 5](#).

Definition 17. (IV Variance Estimator) Define the following variance estimator, which follows [Aronow et al. \(2017\)](#):

$$\begin{aligned} \hat{V}_{n,AS} = & \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{\pi_{n,i}\pi_{n,j}} - \frac{1\{i \neq j\}}{\pi_{n,ij}} \right] \hat{e}_{n,i}(1)\hat{e}_{n,j}(1)D_{n,i}D_{n,j} \\ & + \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{(1-\pi_{n,i})(1-\pi_{n,j})} - \frac{1\{i \neq j\}}{1-\pi_{n,i}-\pi_{n,j}+\pi_{n,ij}} \right] \hat{e}_{n,i}(0)\hat{e}_{n,j}(0)(1-D_{n,i})(1-D_{n,j}) \\ & + 2 \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j \neq i} \frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{\pi_{n,i}(1-\pi_{n,j})(\pi_{n,i} - \pi_{n,ij})} \hat{e}_{n,i}(1)\hat{e}_{n,j}(0)D_{n,i}(1-D_{n,j}) \end{aligned}$$

where

$$\hat{e}_{n,i}(d) = \frac{Y_{n,i} - \hat{\alpha}_{n,IV}W_{n,i} - [\hat{\mu}_{n,Y(d)} - \hat{\alpha}_{n,IV} \hat{\mu}_{n,W(d)}]}{\hat{\mu}_{n,W(1)} - \hat{\mu}_{n,W(0)}}$$

³²[Assumption 6](#) applies to [Definition 16](#) in this case.

This is written entirely in terms of observed outcomes $Y_{n,i}$, observed endogenous regressors $W_{n,i}$, observed treatment indicators $D_{n,i}$, and first- and second-order treatment probabilities. This estimator is consistent for a quantity at least as large as the true variance, as the next proposition states.

Proposition 12. *Assumption 8–Assumption 10,*

$$\hat{V}_{n,AS} - V_n - \frac{1}{N_n} \sum_{i=1}^{N_n} (e_{n,i}(1) - e_{n,i}(0))^2 = o_p(1) \quad (1.7)$$

where V_n is the variance from Proposition *Proposition 11*.

Proof. The proof proceeds in exactly the same way as that of *Proposition 5*. □

1.7 Conclusion

The goal of this paper was to provide an approach for valid inference under spatial assignment mechanisms that induce correlation treatment. By exploiting knowledge of the assignment mechanism (knowledge that is necessary for consistent estimation of causal effects), it is possible to conduct valid inference for the ATE that imposes essentially no assumptions on potential outcomes. I also highlighted that standard use of the *Conley (1999)* estimator leads to overcoverage under correlated assignment, as it fails to exploit knowledge of the assignment mechanism. It underperforms the (modified) variance estimator *Aronow et al. (2017)* in the design-based inference setting. The approach to inference also carried over to the instrumental variables setting, where I derived a conservative variance estimator that can be used to conduct valid inference about a LATE estimand, which is relevant to the applied spatial econometric literature.

Chapter 2

Standard Errors for Calibrated Parameters

2.1 Introduction

Researchers often discipline the parameters of structural economic models by calibrating certain model-implied moments to the corresponding moments in the data (Kydland and Prescott, 1996; Nakamura and Steinsson, 2018). In this paper, we follow Hansen and Heckman (1996) and Thomas Sargent (quoted in Evans and Honkapohja, 2005, p. 568) and take the term “calibration” to be synonymous with moment matching (or more generally, minimum distance) estimation. Moment matching estimators are popular in diverse fields of applied structural economics.

Standard moment matching inference requires knowledge of the entire variance-covariance matrix of the empirical moments, but in practice this matrix is often only partially known. When the empirical moments are obtained from different data sets, different econometric methods, or different previous papers, it is usually hard or impossible to estimate the off-diagonal elements of the variance-covariance matrix. Nevertheless, the diagonal of the matrix—the variances of the individual empirical moments—is typically estimable. In this paper, we show that the diagonal suffices to obtain practically useful worst-case standard errors for the

This chapter is joint work with Mikkel Plagborg-Møller. It has been presented previously at the University of Chicago, New York University, the Federal Reserve Bank of Cleveland, Bocconi University, the North American Summer Meeting of the Econometric Society, the Society for Economic Dynamics Meeting, the Federal Reserve Bank of Philadelphia, MIT, the University of Montreal, Penn State University, the SH3 Conference on Econometrics, Cambridge University, Seoul National University, and the Conference on Computational and Financial Econometrics. This material is also based upon work supported by the NSF under Grant #1851665. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

moment matching estimator. Moreover, in the over-identified case, we show that the moment weighting that minimizes the worst-case estimator variance amounts to a moment selection problem with a simple solution. Hence, our limited-information methods allow researchers to choose their moments and data sources freely without giving up on doing valid statistical inference.

We show that worst-case standard errors for the structural parameters (or smooth transformations thereof), using only the empirical moment variances, are easy to compute. They are given by a weighted sum of the standard errors of individual empirical moments, where the weights depend on the moment weight matrix and the derivatives of the moments with respect to the structural parameters. The derivatives can be obtained analytically, by automatic differentiation, or by first differences. Using these worst-case standard errors, one can construct a confidence interval that is valid even under the worst-case correlation structure. The confidence interval is generally conservative for specific correlation structures, but its minimax coverage is exact, i.e., under the worst-case correlation structure, which amounts to perfect positive/negative correlation. The confidence interval is likely to be informative in many empirical applications, as it is at most $\sqrt{p}$ times wider than it would be if the moments were known to be independent, where p is the number of moments used for estimation.

Given knowledge of only the individual empirical moment variances, we show that the moment weighting scheme that minimizes the worst-case estimator variance amounts to a moment *selection* problem. That is, the efficient minimum distance weight matrix attaches zero weight to some of the moments. The efficient selection of moments can be conveniently computed by running a median regression (i.e., Least Absolute Deviation regression) on a particular artificial data set. Our limited-information efficient estimator given knowledge of only the moment variances is generally different from the familiar full-information efficient estimator that requires knowledge of the entire moment correlation structure.

To understand the intuition behind our efficiency result, consider the analogy of portfolio selection in finance. This analogy is mathematically relevant, as it is well known that any minimum distance estimator is asymptotically equivalent to a linear combination of the empirical moments—a “portfolio” of moments—with a linear restriction on the weights to ensure unbiasedness. When constructing a minimum-variance financial portfolio that achieves a given expected return, it is usually optimal to diversify across all available assets, *except* if the assets are perfectly (positively or negatively) correlated. In the latter extreme case, it is optimal to entirely disregard assets with sufficiently high variance relative to their expected return. But it is precisely the extreme case of perfect correlation that delivers the worst-case variance of a given portfolio. Thus, the portfolio with the smallest worst-case variance across correlation structures is a portfolio that

selects a subset of the available assets.

We propose joint tests of parameter restrictions as well as tests of over-identifying restrictions. A common form of over-identification test used in the empirical literature is to check whether the estimated structural parameters yield a good fit of the model to “non-targeted” moments, i.e., moments that were not exploited for parameter estimation. We show how to implement a formal statistical test based on this idea in our setup. For joint testing of parameter restrictions, we propose a Wald-type test. The proof of the validity of this test relies on tail probability bounds for quadratic forms in Gaussian vectors from Székely and Bakirov (2003), but the test statistic and critical value are simple and easy to compute.

Finally, we extend our limited-information procedures to settings with more detailed knowledge of the covariance matrix of empirical moments. This includes settings where the entire correlation structure is known for some subsets of the moments, or where certain moments are known to be independent of each other.

We illustrate the usefulness of our procedures through two empirical applications. In the first one, we estimate and test the Alvarez and Lippi (2014) model of menu cost price setting in multi-product firms, by matching moments of price changes. In the second application, we estimate and test a heterogeneous agent New Keynesian model developed by McKay, Nakamura and Steinsson (2016) and Auclert, Bardóczy, Rognlie and Straub (2021), by matching impulse responses for macro time series and cross-sectional micro moments. Our worst-case standard errors yield informative inference on several parameters of interest in both applications. A Monte Carlo simulation study calibrated to the first application indicates that our methods perform well in finite samples.

The approach taken in this paper of equating “calibration” with minimum distance estimation is not uncontroversial. Kydland and Prescott (1996) and DeJong and Dave (2011), among others, view “calibration” as a non-econometric procedure for gauging how well a structural macro model fits reduced-form empirical moments. Nevertheless, even in applications where the parameters of the structural model are fixed at values obtained from previous studies, the methods we propose can still be used to compute standard errors for counterfactual quantities and perform tests of the model’s ability to fit additional moments.

Literature. Unlike the literature on correlation matrix completion, we solve the explicit problem of finding the worst-case correlation structure when estimating parameters in a structural model. Existing papers (see Georgescu, Higham and Peters, 2018, and references therein) instead compute positive definite correlation matrices that satisfy various reduced-form optimality criteria, such as the maximum entropy principle. Our

derivations of the worst-case efficient weight matrix and joint testing procedure do not seem to have parallels in the matrix completion literature.

While we focus on cases where it is difficult to estimate the correlation structure of different moments, in some applications it may be possible to model and exploit the precise relationship between the moments. The literature on estimating heterogeneous agent models in macroeconomics has recently developed procedures for combining macro and micro data, as discussed further in [Section 2.5.2](#). [Hahn, Kuersteiner and Mazzocco \(2022a,b\)](#) provide advanced tools for doing inference with a mix of cross-sectional and time series data. These methods, unlike ours, generally require access to the underlying data. [Imbens and Lancaster \(1994\)](#) consider a microeconomic setting where certain macro moments are known without error, which is a special case of our framework. [Hahn, Kuersteiner and Mazzocco \(2020\)](#) give examples of structural models where both time series and cross-sectional data are required for identification of structural parameters. Their insights may help inform the choice of moments for the methods that we develop below.

Outline. [Section 2.2](#) defines the moment matching setup. [Section 2.3](#) derives the worst-case standard errors and the efficient moment weighting/selection. [Section 2.4](#) develops tests of parameter restrictions and of over-identifying restrictions. [Section 2.5](#) contains two empirical illustrations. [Section 2.6](#) concludes. [Appendix B.1](#) contains proofs and other technical details. Code for implementing our procedures is available online.¹

2.2 Setup

Consider a standard moment matching (minimum distance) estimation framework. Let $\mu_0 \in \mathbb{R}^p$ be a vector of reduced-form parameters, which we will refer to as “moments”, though the method applies more generally. Let $\theta_0 \in \Theta \subset \mathbb{R}^k$ be a vector of structural model parameters. According to an economic model, the two parameter vectors are linked by the relationship $\mu_0 = h(\theta_0)$, where $h: \Theta \rightarrow \mathbb{R}^p$ is a known function implied by the model. The map $h(\cdot)$ may be computed either analytically or numerically. We have access to an estimator $\hat{\mu}$ (“empirical moments”) that satisfies

$$\sqrt{n}(\hat{\mu} - \mu_0) \xrightarrow{d} N(0, V) \quad (2.1)$$

¹Matlab: https://github.com/mikkelpm/stderr_calibration_matlab. Python: https://github.com/mikkelpm/stderr_calibration_python

for a $p \times p$ symmetric positive semidefinite variance-covariance matrix V .² Let $\hat{W}$ be a $p \times p$ symmetric matrix satisfying $\hat{W} \xrightarrow{P} W$ (we discuss the choice of $\hat{W}$ in [Section 2.3.2](#)). Then a “moment matching” or “minimum distance” estimator of θ_0 is given by

$$\hat{\theta} = \operatorname{argmin}_{\theta \in \Theta} (\hat{\mu} - h(\theta))' \hat{W} (\hat{\mu} - h(\theta)). \quad (2.2)$$

This estimation strategy is sometimes referred to as “calibration”.³

If we were able to estimate the covariance matrix of the empirical moments $\hat{\mu}$ consistently, it would be straight-forward to construct standard errors for any smooth function of the estimator $\hat{\theta}$. Suppose we are interested in the scalar transformed parameter $r(\theta_0)$, where $r: \Theta \rightarrow \mathbb{R}$. For example, $r(\cdot)$ may equal a particular counterfactual quantity in the structural model, or we could simply set $r(\theta) = \theta_i$ for some index i . Under the standard regularity conditions listed below in [Assumption 11](#),

$$\begin{aligned} \sqrt{n}(r(\hat{\theta}) - r(\theta_0)) &= \lambda'(G'WG)^{-1}G'W\sqrt{n}(\hat{\mu} - \mu_0) + o_p(1) \\ &\xrightarrow{d} N(0, \lambda'(G'WG)^{-1}G'WVWG(G'WG)^{-1}\lambda), \end{aligned} \quad (2.3)$$

where $G \equiv \partial h(\theta_0)/\partial \theta' \in \mathbb{R}^{p \times k}$ and $\lambda \equiv \partial r(\theta_0)/\partial \theta \in \mathbb{R}^k$. See [Newey and McFadden \(1994\)](#) for details. If the entire asymptotic covariance matrix V of $\hat{\mu}$ were consistently estimable, the above display would allow computation of standard errors, confidence intervals, and hypothesis tests.

Unfortunately, the full correlation structure of $\hat{\mu}$ is difficult or impossible to estimate in certain applications. This may be the case, for example, when the moments $\hat{\mu}$ are obtained from a variety of different data sources or econometric methods, or from previous studies for which the underlying data is not readily available. Moreover, if the moments involve a mix of time series and cross-sectional data sources, it can be difficult conceptually or practically to estimate correlations across data sources, whether through the bootstrap or the Generalized Method of Moments (GMM). While the structural model could in some cases be exploited to estimate the moment covariance matrix, this may require stronger assumptions than what

²Here and below, all limits are taken as the sample size $n \rightarrow \infty$. We implicitly think of the sample sizes for the different moments as being proportional, with the factors of proportionality reflected in V . If some element $\hat{\mu}_j$ converges at a faster rate than $\sqrt{n}$, then $V_{jj} = 0$. Sample sizes and convergence rates only enter into our practical procedures through their implicit effect on the calculation of the moment standard errors $\hat{\sigma}_j$ (discussed below), which is handled automatically by econometric software.

³Our analysis extends in a straight-forward manner to Generalized Minimum Distance estimation. In that setting θ_0 and μ_0 are linked through a possibly non-separable equation $g(\theta_0, \mu_0) = 0_{m \times 1}$. The setting in our paper is a special case with $g(\theta, \mu) = h(\theta) - \mu$, but our calculations carry over with few changes because the asymptotic expansions are essentially the same ([Newey and McFadden, 1994](#)).

is needed for point estimation of the structural parameters.⁴ We discuss these points in more detail at the end of this subsection and in the empirical applications in [Section 2.5](#).

Yet, it is often the case that the standard errors of each of the components of $\hat{\mu}$ are available. These marginal standard errors may be directly computable from data, or they may be reported in the various papers that the individual elements of $\hat{\mu}$ are gleaned from. Thus, assume that we have access to standard errors $\hat{\sigma}_1, \dots, \hat{\sigma}_p \geq 0$ satisfying

$$\sqrt{n}\hat{\sigma}_j \xrightarrow{p} V_{jj}^{1/2}, \quad j = 1, \dots, p. \quad (2.4)$$

We show in the next section that these marginal standard errors suffice for doing informative inference on $r(\theta_0)$.⁵

For ease of reference, we summarize our technical assumptions here:

Assumption 11.

- i) The empirical moment vector $\hat{\mu}$ is asymptotically normal, as in [\(2.1\)](#).
- ii) The standard errors $\hat{\sigma}_j$ are consistent, as in [\(2.4\)](#).
- iii) $h(\cdot)$ and $r(\cdot)$ are both continuously differentiable in a neighborhood of θ_0 (which lies in the interior of Θ), $G \equiv \partial h(\theta_0)/\partial \theta'$ has full column rank, and $\lambda \equiv \partial r(\theta_0)/\partial \theta \neq 0_{k \times 1}$.
- iv) $\hat{\theta} \xrightarrow{p} \theta_0$.
- v) $\hat{W} \xrightarrow{p} W$ for a symmetric positive semidefinite matrix W .
- vi) $G'WG$ is nonsingular.

Conditions (ii)–(vi) are standard regularity conditions that are satisfied in smooth, identified models ([Newey and McFadden, 1994](#)). Note that we allow for the possibility that some moments are known with certainty ($V_{jj} = 0$) as in [Imbens and Lancaster \(1994\)](#). We will now discuss the key condition (i).

Discussion of limited information about the correlation structure and the joint normality assumption. We now provide several examples of situations where it is difficult or impossible to estimate the correlation structure of different types of empirical moments. In cases when the elements of the moment

⁴For example, if we exploit the model's predictions about second moments for estimation, model-based estimation of V would require believing the model's predictions about fourth moments.

⁵Our set-up covers applications where the parameters $\hat{\theta}$ are not explicitly estimated but are instead fixed at certain values $\hat{\mu}$ obtained from prior studies. In this case $h(\theta) = \theta$, and $\hat{\sigma}_j$ is the standard error (which could potentially be 0 for some j) of the j -th externally-obtained parameter value $\hat{\theta}_j = \hat{\mu}_j$.

vector $\hat{\mu}$ are obtained from different data sets, the joint normality assumption (2.1) requires justification. This ensures not only that a normal distribution is the appropriate reference distribution for obtaining critical values, but also that the vector $\hat{\mu}$ can reasonably be viewed as arising from some joint, repeatable experiment for which the given standard errors $\hat{\sigma}_j$ capture all sources of uncertainty. The joint normality assumption is most easily understood and justified under a model-based (e.g., shock-based) perspective on uncertainty. In this view, there exists a coherent data generating process with both aggregate and idiosyncratic shocks that affect all of the observed data. The empirical application of Section 2.5.2 is a prototypical example of this framework, but such applications are not the only use case.

There are several cases in which the joint normality assumption appears reasonable, but estimation of the full moment covariance matrix V could be challenging. For example:

1. *The moments are obtained from the same or similar data sets, but the underlying data for some of the moments is not available.* For example, some moments may be reported in previous papers that use proprietary data. In this case, traditional full-information inference procedures are inapplicable because it is generally impossible to estimate moment cross-correlations without access to the underlying data. However, as long as we have access to marginal standard errors for each individual moment, our methods can be applied. See Section 2.5.1 for an empirical example.
2. *Some of the moments are computed from aggregate time series and others from panel data spanning similar time periods.* If the clustering procedure of the panel data regressions allows for aggregate shocks (such as when clustering by time period or when using the standard error formula from Driscoll and Kraay, 1998), and these aggregate shocks also affect the time series data, then the panel regressions will have correct standard errors but the coefficients may be correlated with the time series moments.
3. *The moments stem from time series data observed at various frequencies, or from regional data with various levels of geographic aggregation.* While careful econometric analysis may allow the estimation of the full covariance matrix of the moments by appropriately collapsing the sample moment conditions for the higher-frequency or higher-resolution data, this could be cumbersome in practice. This is especially true if there is limited overlap in the time spans of the higher-frequency and lower-frequency data, or in the geographical coverage of the higher-resolution and lower-resolution data.
4. *We use a combination of aggregate time series moments and micro moments from surveys, and the latter measure time-invariant parameters that are not affected by macro shocks in the sample.* In this case, it is often reasonable to assume that the uncertainty in the micro moments (arising purely

from idiosyncratic noise) is uncorrelated with the uncertainty in the macro moments. Similarly, cross-sectional moments obtained from different independent random samples are plausibly uncorrelated with each other. Such extra information about off-diagonal entries of V can be incorporated in our procedures, as shown in [Section 2.3.3](#).

5. *The moments are all computed from the same data set, but using a variety of complicated procedures.*

While it may in principle be possible to estimate the correlation structure analytically by stacking the first-order conditions for the different estimation procedures into a large GMM moment vector, applied researchers may wish to avoid these complications. The bootstrap offers a simpler approach but is computationally impractical in some cases. In contrast, our analytical inference procedures below only require researchers to obtain standard errors for each estimated moment separately, and these are often outputted automatically by existing software. One may worry about the internal consistency of mixing moments arising from different estimation procedures, in the sense that the models used to compute the various moments could be mutually incompatible (or incompatible with the structural model that the researcher seeks to fit to the moments). This can be tested using the over-identification test that we propose in [Section 2.4.2](#).

6. *The moments are all computed from the same data set, but full-information inference procedures perform poorly in finite samples.* As shown by [Altonji and Segal \(1996\)](#), traditional full-information efficient minimum distance estimation, which relies on an estimate $\hat{V}$ of the entire moment variance-covariance matrix V , can be subject to severe finite-sample biases in many applications, especially when p is large.⁶ Whereas full-information efficient inference requires the entire matrix V to be accurately estimated, our limited-information inference procedures below only require accurate estimates of the diagonal entries.

To be clear, in cases 2–6 above (though not in case 1), full-information inference may be possible in principle, as we have indicated. However, this could be cumbersome to implement in practice and/or suffer from poor finite-sample properties. The limited-information procedures we develop offer applied researchers simpler tools that in many cases provide informative inference about structural parameters. Moreover, the limited-information results could be used as a first step for exploring whether it is worthwhile to expend the extra effort required for full-information inference.

⁶When matching simple moments with access to the underlying i.i.d. data, generalized empirical likelihood estimators are known to have better finite-sample properties than efficient minimum distance or GMM, though they are often harder to compute (see [Imbens, 2002](#), for a review).

It is important to note that the joint normality assumption (2.1) may fail in some cases, rendering our methods invalid. For example:

1. *We use a combination of aggregate time series moments and micro moments from surveys, but the latter are affected by aggregate macro shocks that shift the whole micro outcome distribution.* In this case, standard cross-sectional moments may not even be *consistent* for the true underlying population moments, since the aggregate shocks do not get averaged out (Hahn et al., 2020, Section 3). Moreover, the usual micro standard errors will not take into account the *combined* uncertainty in the macro shocks and idiosyncratic micro noise. However, if the effects of macro shocks on micro moments do average out and we appropriately account for all types of uncertainty when computing moments and their standard errors, then our methods below can be applied. We illustrate this empirically in Section 2.5.2.
2. *The data used to compute some of the moments is very heavy tailed, or the estimation procedures are not asymptotically regular.* In this case, even *marginal* normality of the individual empirical moments may fail. We leave extensions to non-normal limit distributions as an interesting topic for future research.

2.3 Standard errors and moment selection

We first derive the worst-case standard errors for a given choice of moment weight matrix. Then we show that the weighting scheme that minimizes the worst-case standard errors amounts to a moment selection problem with a simple solution.

2.3.1 Worst-case standard errors and confidence intervals

We first compute the worst-case bound on the standard error of the moment matching estimator, given knowledge of only the variances of the empirical moments. Although the mathematical argument is straightforward, it appears that the literature has not realized the practical utility of this result.

Recall that we seek to do inference on the scalar parameter $r(\theta_0)$. By the standard delta method expansion (2.3) under Assumption 11, the estimator $r(\hat{\theta})$ is asymptotically equivalent to a certain linear function $x'\hat{\mu}$ of the empirical moments, where $x = (x_1, \dots, x_p)'\equiv WG(G'WG)^{-1}\lambda$. We thus seek to bound the variance of an (asymptotically) known linear combination of $\hat{\mu}$, knowing the variance of each component $\hat{\mu}_j$ but not the correlation structure. This worst-case variance is attained when all components of $\hat{\mu}$ are perfectly positively

or negatively correlated (depending on the signs of the elements of x), yielding the worst-case variance $(\sum_{j=1}^p |x_j| \text{Var}(\hat{\mu}_j)^{1/2})^2$, as is proved in the following elementary lemma.⁷

Lemma 2. *Let $x = (x_1, \dots, x_p)' \in \mathbb{R}^p$ and $\sigma_1^2, \dots, \sigma_p^2 \geq 0$. Let $\mathcal{S}(\sigma)$ denote the set of $p \times p$ symmetric positive semidefinite matrices with diagonal elements $\sigma = (\sigma_1^2, \dots, \sigma_p^2)'$. Then*

$$\max_{V \in \mathcal{S}(\sigma)} \sqrt{x' V x} = \sum_{j=1}^p \sigma_j |x_j|.$$

Proof. The right-hand side is attained by $V = ss'$, where $s = (\sigma_1 \text{sign}(x_1), \dots, \sigma_p \text{sign}(x_p))'$. Moreover, for any $V \in \mathcal{S}(\sigma)$,

$$x' V x = \sum_{j=1}^p \sum_{\ell=1}^p x_j x_\ell V_{j\ell} \leq \sum_{j=1}^p \sum_{\ell=1}^p |x_j x_\ell| |V_{j\ell}| \leq \sum_{j=1}^p \sum_{\ell=1}^p |x_j x_\ell| \sigma_j \sigma_\ell = \left(\sum_{j=1}^p \sigma_j |x_j| \right)^2.$$

The penultimate inequality uses that $|V_{j\ell}|^2 \leq V_{jj} V_{\ell\ell}$ for any symmetric positive semidefinite matrix V . $\square$

We can thus construct an estimate of the worst-case standard error of $r(\hat{\theta})$ as

$$\widehat{\text{se}}(\hat{x}) \equiv \sum_{j=1}^p \hat{\sigma}_j |\hat{x}_j|, \quad (2.5)$$

where $\hat{x} = (\hat{x}_1, \dots, \hat{x}_p)' \equiv \hat{W} \hat{G} (\hat{G}' \hat{W} \hat{G})^{-1} \hat{\lambda}$, $\hat{G} \equiv \frac{\partial h(\hat{\theta})}{\partial \theta'}$, and $\hat{\lambda} \equiv \frac{\partial r(\hat{\theta})}{\partial \theta}$. In practice, the derivatives may be computed analytically, by automatic differentiation, or by finite differences. Let $\Phi(\cdot)$ denote the standard normal distribution function. Then the confidence interval

$$\widehat{\text{CI}} \equiv \left[r(\hat{\theta}) - \Phi^{-1}(1 - \alpha/2) \widehat{\text{se}}(\hat{x}), r(\hat{\theta}) + \Phi^{-1}(1 - \alpha/2) \widehat{\text{se}}(\hat{x}) \right]$$

covers $r(\theta_0)$ with probability at least $1 - \alpha$ asymptotically, as shown formally in the following proposition. The asymptotic coverage probability generally exceeds $1 - \alpha$ due to the worst-case perspective, but coverage is exact in a particular special case when all elements of the empirical moment vector $\hat{\mu}$ are perfectly correlated asymptotically.

Proposition 13. *Impose [Assumption 11](#) and $\max_j |x_j| V_{jj} > 0$. Then $\lim_{n \rightarrow \infty} P(r(\theta_0) \in \widehat{\text{CI}}) \geq 1 - \alpha$, and the inequality binds for the rank-1 matrix V defined in the proof of [Lemma 2](#).*

⁷The basic intuition is that $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2[\text{Var}(X) \text{Var}(Y)]^{1/2} \text{Corr}(X, Y) \leq \text{Var}(X) + \text{Var}(Y) + 2[\text{Var}(X) \text{Var}(Y)]^{1/2} = (\text{Var}(X)^{1/2} + \text{Var}(Y)^{1/2})^2$, since $\text{Corr}(X, Y) \leq 1$.

Proof. Please see [Appendix B.1.1](#). □

Remarks.

1. The worst-case standard errors are at most $\sqrt{p}$ times larger than the standard errors that assume all elements of $\hat{\mu}$ to be mutually uncorrelated. This follows from Hölder's inequality between ℓ_1 and ℓ_2 norms, which yields $\sum_{j=1}^p \hat{\sigma}_j |\hat{x}_j| \leq p^{1/2} (\sum_{j=1}^p \hat{\sigma}_j^2 \hat{x}_j^2)^{1/2}$.
2. Though often informative, limited-information inference can potentially have much lower power than full-information inference. Specifically, the worst-case standard errors can in some (but not all) models be arbitrarily larger than the standard errors one would report given full knowledge of V . For example, consider a “repeated measurements” model with $k = 1$, $p = 2$, $h(\theta) = (\theta, \theta)'$, $r(\theta) = \theta$, and $\hat{W} = I$, yielding the estimator $\hat{\theta} = (\hat{\mu}_1 + \hat{\mu}_2)/2$. The worst-case standard error of $\hat{\theta}$ equals $(\hat{\sigma}_1 + \hat{\sigma}_2)/2$, corresponding to the case where $\hat{\mu}_1$ and $\hat{\mu}_2$ are perfectly positively correlated. But it is easy to verify that the *best-case* standard error equals $|\hat{\sigma}_1 - \hat{\sigma}_2|/2$, corresponding to $\hat{\mu}_1$ and $\hat{\mu}_2$ being perfectly negatively correlated. If $\hat{\sigma}_1 = \hat{\sigma}_2$, the ratio of worst-case to best-case standard errors is infinite.
3. In applications where the moment function $h(\theta)$ cannot be evaluated analytically, the moments can instead be approximated by simulating from the economic model given parameter vector θ . [Proposition 21](#) in [Appendix B.1.2](#) shows that the conclusions of [Proposition 13](#) continue to apply in this case, as long as the number of simulation draws is sufficiently large relative to the empirical sample size n .

2.3.2 Efficient moment selection

We now derive a weight matrix W that minimizes the *worst-case* variance of the estimator, derived above. We show that this weight matrix puts weight on at most k moments, so the procedure amounts to efficient moment *selection*. Since the weight matrix W only matters in the over-identified case, we assume $p > k$ in this section.

We seek a weight matrix W that minimizes the worst-case asymptotic standard deviation of $r(\hat{\theta})$. Let $x(W)$ denote the vector x defined in [Section 2.3.1](#), viewed as a function of W , and let $\mathcal{S}_p$ denote the set of $p \times p$ symmetric positive semidefinite matrices W such that $G'WG$ is nonsingular. Then we solve the

problem

$$\begin{aligned}
& \min_{W \in \mathcal{S}_p} \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} \sqrt{\lambda'(G'WG)^{-1}G'W\tilde{V}WG(G'WG)^{-1}\lambda} \\
&= \min_{W \in \mathcal{S}_p} \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} (x(W)' \tilde{V} x(W))^{1/2} \\
&= \min_{W \in \mathcal{S}_p} \sum_{j=1}^p V_{jj}^{1/2} |x_j(W)|,
\end{aligned} \tag{2.6}$$

where the last equality uses [Lemma 2](#) in [Section 2.3.1](#). [Lemma 11](#) in [Appendix B.1.1](#) shows that the solution to the final optimization problem above is given by

$$\min_{W \in \mathcal{S}_p} \sum_{j=1}^p V_{jj}^{1/2} |x_j(W)| = \min_{z \in \mathbb{R}^{p-k}} \sum_{j=1}^p |\tilde{Y}_j - \tilde{X}_j' z|, \tag{2.7}$$

where we define

$$\tilde{Y}_j \equiv V_{jj}^{1/2} G_{j\bullet} (G'G)^{-1} \lambda \in \mathbb{R}, \quad \tilde{X}_j \equiv -V_{jj}^{1/2} G_{j\bullet}^{\perp} \in \mathbb{R}^{p-k},$$

$G^{\perp}$ is any $p \times (p-k)$ matrix with full column rank satisfying $G'G^{\perp} = 0_{k \times (p-k)}$, and the notation $A_{j\bullet}$ means the j -th row of matrix A . The intuition for the equality (2.7) is that the set of all minimum distance estimators for various weight matrices is asymptotically equivalent to the set of all estimators that are linear combinations of the p moments $\hat{\mu}$, subject to k asymptotic unbiasedness constraints. We can therefore optimize over a $(p-k)$ -dimensional linear space.

The final optimization problem (2.7) is a median regression (Least Absolute Deviation regression) of the artificial “regressand” $\{\tilde{Y}_j\}$ on the $p-k$ artificial “regressors” $\{\tilde{X}_j\}$. This regression can be executed efficiently using standard quantile regression software.

The solution to the median regression amounts to optimally selecting at most k of the p moments for estimation. Theorem 3.1 of [Koenker and Bassett \(1978\)](#) implies that there exists a solution z^* to the median regression (2.7) such that at least $p-k$ out of the p median regression residuals

$$e_j^* \equiv \tilde{Y}_j - \tilde{X}_j' z^*, \quad j = 1, \dots, p,$$

equal zero. Hence, an efficient weight matrix W^* that achieves the minimum in (2.7) will yield a linear combination vector $x(W^*) = (V_{11}^{-1/2} e_1^*, \dots, V_{pp}^{-1/2} e_p^*)'$ that attaches nonzero weight to at most k out of the p empirical moments $\hat{\mu}$. In other words, the solution to the efficient moment *weighting* problem is endogenously

achieved by an efficient moment *selection*. We may pick an arbitrary weight matrix that attaches nonzero weight to only the efficiently selected moments (the magnitudes of the weights do not matter asymptotically, as the selected set of moments is just-identified).

Analytical illustration. To illustrate the above results, and to link back to the intuitive discussion of portfolio selection in [Section 2.1](#), consider a “repeated measurements” model with $k = 1, p = 2, h(\theta) = (\theta, \theta)'$, and $r(\theta) = \theta$. It is easy to see that any weight matrix $\hat{W}$ yields a minimum distance estimator of the form $\hat{\theta} = x_1 \hat{\mu}_1 + x_2 \hat{\mu}_2$, where $x_1 + x_2 = 1$. The constraint that x_1 and x_2 sum to 1 ensures that the estimator is consistent. The choice of optimal weight matrix corresponds to an optimal choice of x_1 and $x_2 = 1 - x_1$. One might expect the usual diversification motive to cause us to attach nonzero weights to both empirical moments $\hat{\mu}_1$ and $\hat{\mu}_2$, so that their estimation errors average out partially. However, for the worst-case correlation structure for these moments—perfect correlation—their estimation errors amplify each other, rendering diversification futile. Hence, we minimize the worst-case standard error $|x_1| \hat{\sigma}_1 + |1 - x_1| \hat{\sigma}_2$ by setting $x_1 = 1$ when $\hat{\sigma}_1 < \hat{\sigma}_2$ and otherwise $x_1 = 0$, yielding the estimator $\hat{\theta} = \hat{\mu}_{\arg\min_j \hat{\sigma}_j}$. That is, the efficient estimator simply discards the moment with the largest standard error. In more general models, moment selection also depends on how informative each moment is about the structural parameters, as measured by the Jacobian matrix G .

Algorithm. The efficient estimator and standard errors can be computed as follows:

- i) Compute an initial $\sqrt{n}$ -consistent estimator $\hat{\theta}_{\text{init}}$ using, say, a diagonal weight matrix with $\hat{W}_{jj} = \hat{\sigma}_j^{-2}$.
- ii) Construct the derivative matrix $\hat{G} \equiv \frac{\partial h(\hat{\theta}_{\text{init}})}{\partial \theta'}$ and vector $\hat{\lambda} \equiv \frac{\partial r(\hat{\theta}_{\text{init}})}{\partial \theta}$, either analytically or numerically.
- iii) Solve the median regression (2.7), substituting $\hat{G}$ for G , $\hat{\lambda}$ for λ , and $\hat{\sigma}_j$ for $V_{jj}^{1/2}$.⁸ Compute the residuals $\hat{e}_j^*, j = 1, \dots, p$, from this median regression. (In the non-generic case where multiple solutions to the median regression exist, select one that yields at most k nonzero residuals.)
- iv) Construct the efficient linear combination $\hat{x}^* = (\hat{x}_1^*, \dots, \hat{x}_p^*)'$ of the p moments, given by $\hat{x}_j^* \equiv \hat{\sigma}_j^{-1} \hat{e}_j^*$ for $j = 1, \dots, p$. At least $p - k$ of the elements will be zero, corresponding to those moments that are discarded by the efficient moment selection procedure.

- v) To compute an efficient estimator of $r(\theta_0)$, either:

⁸Remember to omit an intercept from the regression.

- a) Compute the just-identified efficient minimum distance estimator $\hat{\theta}_{\text{eff}}$ of θ_0 which uses any weight matrix that attaches zero weight to those (at least) $p - k$ moments which receive zero weight in the vector $\hat{x}^*$. Then estimate $r(\theta_0)$ by $r(\hat{\theta}_{\text{eff}})$. Or:
- b) Compute the “one-step” estimator $\hat{r}_{\text{eff-1S}} \equiv r(\hat{\theta}_{\text{init}}) + \hat{x}^{*'}(\hat{\mu} - h(\hat{\theta}_{\text{init}}))$ of $r(\theta_0)$.⁹
- vi) The worst-case standard error of the estimator from step (v)) is given by the value of the median regression (2.7) (i.e., the minimized objective function).

Options (a) and (b) in step (v)) of the algorithm are asymptotically equivalent. Option (b) is computationally more convenient as it avoids further numerical optimization, but option (a) ensures that $\hat{\theta}_{\text{eff}}$ always lies in the parameter space Θ .

Remarks.

1. Since all operations involved in computing the efficient linear combination $\hat{x}^*$ are continuous, $\hat{x}^*$ converges in probability to the population efficient linear combination $x(W^*)$. The only exception may be where the population median regression (2.7) does not have a unique minimum, which is a non-generic case. Even in this case, however, the efficient worst-case standard errors will be consistent (when multiplied by $\sqrt{n}$) under Assumption 11(i)–(iv)), by a standard application of the maximum theorem.
2. The full-information (infeasible) efficient weight matrix that exploits knowledge of all of V is known to equal $W = V^{-1}$. This weight matrix in general attaches nonzero weight to all moments, unlike the limited-information efficient solution derived above. The worst-case asymptotic standard deviation (2.7) given limited information is of course larger than the asymptotic standard deviation $(\lambda'(G'V^{-1}G)^{-1}\lambda)^{1/2}$ of the full-information efficient estimator of $r(\theta_0)$.
3. The efficient moment weighting/selection for the limited-information efficient estimators $r(\hat{\theta}_{\text{eff}})$ and $\hat{r}_{\text{eff-1S}}$ depends on the function $r(\cdot)$ of interest, unlike in the case of full-information efficient estimation. In practice, we can just re-run the computations for each function $r(\cdot)$ of interest (e.g., for each component of θ).
4. By restricting W to the set $\mathcal{S}_p$ of weight matrices for which $G'WG$ is nonsingular, we ensure that the parameter vector θ_0 is at least *locally* identified from the selected moments (Newey and McFadden, 1994, p. 2144). However, the selected moments may fail to ensure *global* identification of θ_0 . In models

⁹See Newey and McFadden (1994, Section 3.4) for a general discussion of one-step estimators.

for which this is a concern, we recommend using the one-step efficient estimator $\hat{r}_{\text{eff-1S}}$, which inherits its consistency from the initial estimator $\hat{\theta}_{\text{init}}$ that could be computed using all available moments.

5. It is not restrictive to consider moment matching estimators of the form (2.2). Consider instead any estimator $\hat{\vartheta} \equiv \hat{f}(\hat{\mu})$ of θ_0 , where $\hat{f}: \mathbb{R}^p \rightarrow \mathbb{R}^k$ is a possibly data-dependent function with enough regularity to satisfy the asymptotically linear expansion

$$\hat{\vartheta} - \theta_0 = H(\hat{\mu} - \mu_0) + o_p(n^{-1/2}),$$

for some $k \times p$ matrix H . If we restrict attention to asymptotically regular estimators of θ_0 (i.e., estimators that remain asymptotically unbiased under locally drifting parameters), we need $HG = I_k$. Among all estimators $\hat{\vartheta}$ satisfying these requirements, the smallest possible worst-case asymptotic standard deviation of $r(\hat{\vartheta})$ is achieved by the estimator whose asymptotic linearization matrix H solves

$$\min_{H: HG=I_k} \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} (\lambda' H' \tilde{V} H \lambda)^{1/2}.$$

Lemma 11 in Appendix B.1.1 shows that the solution to this problem is precisely the value of the median regression (2.7). In other words, the minimum distance estimator $\hat{\theta}$ with (limited-information) efficient weight matrix delivers the smallest possible worst-case standard errors in a large class of estimators.

2.3.3 General knowledge of the covariance matrix

We now extend our results to the general case where any collection of elements of the asymptotic covariance matrix V of the moments $\hat{\mu}$ is known (or consistently estimable), while the remaining elements are unrestricted. For example, if a pair of elements of $\hat{\mu}$ are known to be independent, the corresponding off-diagonal elements of V must equal zero.

Letting $\tilde{\mathcal{S}}$ denote the given constraint set for V , we can compute the worst-case asymptotic standard deviation of $r(\hat{\theta})$ as

$$\sqrt{\max_{\tilde{V} \in \tilde{\mathcal{S}}} x' \tilde{V} x}, \tag{2.8}$$

where x was defined in Section 2.3.1.¹⁰ In the case of interest to us, $\tilde{\mathcal{S}}$ is defined by equality restrictions on a subset of the elements of V , in addition to the restriction that V is symmetric positive semidefinite.¹¹

¹⁰Similarly, we could compute the *best-case* standard deviation by minimizing this objective function.

¹¹In fact, the argument also extends to imposing *inequality* constraints on elements of V .

Then the maximization problem (2.8) is a so-called semidefinite programming problem, a special case of convex programming. In [Appendix B.1.3](#), we show that the efficient weight matrix can likewise be computed through a pair of nested convex optimizations. We also show that analytical simplifications obtain in the special case where we know the *block* diagonal of V ; such a structure may occur if consecutive elements of $\hat{\mu}$ are obtained from the same data set, allowing estimation of covariances among these elements. Our software automatically implement all these computations (see [Footnote 1](#)).

2.4 Testing

In this section we develop a joint test of multiple parameter restrictions as well as a test of over-identifying restrictions.

2.4.1 Joint testing of parameter restrictions

We propose a test of the joint null hypothesis $H_0: r(\theta_0) = 0_{m \times 1}$ against the two-sided alternative. In this section, the continuously differentiable function $r: \Theta \rightarrow \mathbb{R}^m$ is allowed to be multi-valued. Tests of a single parameter restriction ($m = 1$) can be carried out using the confidence interval described in [Section 2.3.1](#). For the case $m > 1$, we propose the following testing procedure. Let $\alpha \in (0, 1)$ denote the significance level.

- i) Compute the Wald-type test statistic

$$\hat{\mathcal{T}} \equiv r(\hat{\theta})' \hat{S} r(\hat{\theta}),$$

where $\hat{S}$ is a user-specified symmetric positive definite $m \times m$ matrix, to be discussed below.

- ii) Compute the critical value

$$cv_n \equiv \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} \frac{1}{n} \text{trace} \left(\tilde{V} W G (G' W G)^{-1} \lambda S \lambda' (G' W G)^{-1} G' W \right) \times \left(\Phi^{-1}(1 - \alpha/2) \right)^2. \quad (2.9)$$

In practice, we substitute the estimates $\frac{1}{n} \text{diag}(V) \approx \text{diag}(\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2)$, $W \approx \hat{W}$, $S \approx \hat{S}$, $G \approx \hat{G} \equiv \frac{\partial h(\hat{\theta})}{\partial \theta'}$, and $\lambda \approx \hat{\lambda} \equiv \frac{\partial r(\hat{\theta})'}{\partial \theta}$.

- iii) Reject $H_0: r(\theta_0) = 0_{m \times 1}$ if $\hat{\mathcal{T}} > cv_n$.

The maximization (2.9) is a semidefinite programming problem, which is easy to compute numerically, as discussed in [Section 2.3.3](#). [Appendix B.1.3](#) extends the procedure to settings where additional knowledge about V is available other than the diagonal.

The following proposition shows that, for conventional significance levels α , the asymptotic size of this test does not exceed α , regardless of the true correlation structure of the moments. This holds for any valid choice of weight matrix $\hat{W}$, including—but not limited to—the limited-information efficient weight matrix derived in [Section 2.3.2](#).

Proposition 14. *Impose [Assumption 11](#), except that we redefine $\lambda \equiv \partial r(\theta_0)' / \partial \theta$ and require this matrix to have full column rank m . Assume also that $\hat{S} \xrightarrow{P} S$, S is symmetric positive definite, $n \text{cv}_n > 0$, and $\alpha \leq 0.215$. Then, if $r(\theta_0) = 0_{m \times 1}$,*

$$\limsup_{n \rightarrow \infty} P(\hat{\mathcal{T}} > \text{cv}_n) \leq \alpha.$$

Proof. Please see [Appendix B.1.1](#). □

Remarks.

1. We do not have formal results on how to choose the weight matrix $\hat{S}$ in the test statistic. A pragmatic *ad hoc* choice is to set $\hat{S} = (\lambda'(G'WG)^{-1}G'W\bar{V}WG(G'WG)^{-1}\lambda)^{-1}$ (with consistent estimates plugged in), where $\bar{V}$ is a particular value for the asymptotic variance V of $\hat{\mu}$ at which we wish to direct power. Then $\hat{\mathcal{T}}$ is the usual full-information Wald statistic given $V = \bar{V}$, though the critical value for the test differs from the conventional one to ensure size control regardless of the actual unknown correlation structure. For example, if we want to direct power toward the case with independent moments, we can choose $\bar{V} = \text{diag}(n\hat{\sigma}_1^2, \dots, n\hat{\sigma}_p^2)$.
2. The test procedure in [Proposition 14](#) is generally conservative from a minimax perspective, i.e., the size may be strictly smaller than α for all covariance matrices V of the moments. The reason is that the proof of [Proposition 14](#) relies on a tail probability bound for quadratic forms of Gaussian random vectors ([Székely and Bakirov, 2003](#)). This bound is attained when V has rank 1, but the positive semidefinite maximum [\(2.9\)](#) need not be attained at a rank-1 matrix, to our knowledge. It is an interesting topic for future research to devise a test that has a formal minimax optimality property given the limited knowledge of V .
3. The test procedure is consistent against any fixed alternative with $r(\theta_0) \neq 0_{m \times 1}$ under the conditions of [Proposition 14](#). This follows from the standard argument that $n\hat{\mathcal{T}}$ diverges to infinity with probability 1 in this case, while $\text{cv}_n = O(n^{-1})$ since the largest eigenvalue of any matrix $\tilde{V} \in \mathcal{S}(\text{diag}(V))$ is bounded above by $\sum_{j=1}^p V_{jj}$.

4. An alternative way to test multiple hypotheses ($m > 1$) is to separately test each null hypothesis $r_j(\theta_0) = 0$, $j = 1, \dots, m$, using the univariate confidence interval for $r_j(\theta_0)$ in [Section 2.3.1](#), but with a Bonferroni correction applied to the significance level.¹² That is, the test rejects the joint hypothesis when any of the m level- $(1 - \alpha/m)$ confidence intervals excludes 0. This Bonferroni test neither dominates, nor is it dominated by, the test in [Proposition 14](#) in terms of local power in general. The latter test has the advantage that it is finite-sample invariant to nonsingular linear transformations of the joint null hypothesis (if $\hat{S}$ is chosen as in the first remark above).

2.4.2 Over-identification testing

The fit of the calibrated model can be evaluated using over-identification tests when we have more moments p than parameters k . In this subsection we allow for potential model misspecification by dropping the assumption in [Section 2.2](#) that there exists $\theta_0 \in \mathbb{R}^k$ such that $h(\theta_0) = \mu_0$. Let an arbitrary weight matrix $\hat{W} \xrightarrow{P} W$ be given, such as the limited-information efficient weight matrix derived in [Section 2.3.2](#). Define the pseudo-true parameter $\tilde{\theta}_0 \equiv \operatorname{argmin}_{\theta \in \mathbb{R}^k} (\mu_0 - h(\theta))' W (\mu_0 - h(\theta))$, assuming the minimizer is unique. We continue to impose all the assumptions in [Section 2.2](#), with $\tilde{\theta}_0$ substituting for θ_0 .

Suppose we want to know whether the model provides a good fit for a particular moment. Let $j^* \in \{1, \dots, p\}$ be the index of the moment of interest. We seek a confidence interval for the model misspecification measure $\mu_{0,j^*} - h_{j^*}(\tilde{\theta}_0)$, i.e., the j^* -th element of $\mu_0 - h(\tilde{\theta}_0)$. It is standard to show that, under [Assumption 11](#),

$$\hat{\mu} - h(\hat{\theta}) - (\mu_0 - h(\tilde{\theta}_0)) = (I_p - G(G'WG)^{-1}G'W)(\hat{\mu} - \mu_0) + o_p(n^{-1/2}). \quad (2.10)$$

Let $\bar{x}$ be the j^* -th column of the matrix $I_p - \hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'$. Then

$$\left[\hat{\mu}_{j^*} - h_{j^*}(\hat{\theta}) - \Phi^{-1}(1 - \alpha/2)\widehat{\text{se}}(\bar{x}), \hat{\mu}_{j^*} - h_{j^*}(\hat{\theta}) + \Phi^{-1}(1 - \alpha/2)\widehat{\text{se}}(\bar{x}) \right]$$

is a confidence interval for the difference $\mu_{0,j^*} - h_{j^*}(\tilde{\theta}_0)$, with worst-case asymptotic coverage probability $1 - \alpha$. Note that it can happen that $\widehat{\text{se}}(\bar{x}) = 0$, in which case it is not possible to test the over-identifying restriction corresponding to the j^* -th moment.¹³

¹²We thank a referee for this suggestion. Yet another approach was suggested to us by Bo Honoré: Reject whenever $\inf_V nr(\hat{\theta})'(\lambda'(G'WG)^{-1}G'WVWG(G'WG)^{-1}\lambda)^{-1}r(\hat{\theta})$ exceeds the $1 - \alpha$ quantile of a chi-squared distribution with m degrees of freedom. That is, we search over V for the smallest conventional Wald test statistic. Unfortunately, this optimization problem appears to be numerically challenging unless p is small.

¹³Similarly, following the ideas in [Section 2.4.1](#), a *joint* test of the over-identifying restrictions can be constructed by computing the test statistic $\hat{\mathcal{T}}_{\text{overid}} \equiv (\hat{\mu} - h(\hat{\theta}))'\hat{S}(\hat{\mu} - h(\hat{\theta}))$ for some $p \times p$ symmetric positive definite matrix $\hat{S}$. A natural *ad hoc* choice is $\hat{S} = \hat{W}$, in which case the test statistic equals the minimized minimum distance objective function. We reject

One common use of over-identification testing is to evaluate the estimated model’s fit on “non-targeted moments”. This corresponds to the special case where the weight matrix $\hat{W}$ zeroes out the corresponding rows and columns of the non-targeted moments, so that the point estimate $\hat{\theta}$ ignores these moments.¹⁴ An application of this idea is to fix the structural parameters θ to values obtained from previous studies and then evaluate whether the model’s predictions at these parameters match additional moments (this can be viewed as a statistical version of the model validation process advocated by [Kydland and Prescott, 1996](#)). This is achieved by choosing a moment function of the form $h(\theta) = (\theta', \bar{h}(\theta))'$, with empirical moments $\hat{\mu} = (\hat{\theta}', \hat{\mu})'$. Here $\hat{\theta}$ are the parameter values obtained from previous studies, while $\bar{h}(\theta)$ is the “non-targeted” model prediction of the empirical moment $\hat{\mu}$.

2.5 Empirical applications

We illustrate our methods through two empirical examples. First we fit a model of menu cost price setting in multi-product firms to scanner data. Then we fit a heterogeneous agent New Keynesian model to impulse responses that have been estimated from a combination of micro and macro data.

2.5.1 Menu cost price setting in multi-product firms

Our first application estimates the [Alvarez and Lippi \(2014\)](#) model of menu cost price-setting in multiproduct firms. We fit the model to moments of price changes from supermarket scanner data. This is a small-scale application with $k = 3$ parameters and $p = 4$ moments.

Though we in fact have access to the underlying data set, we emulate a hypothetical situation where the model is matched to moments that were reported in another paper. We can therefore compare full-information inference, which uses the underlying data, with limited-information inference, which uses only the moments and their marginal standard errors. We find that limited-information inference remains informative about the structural parameters. Moreover, a simulation study calibrated to this application confirms the utility of our procedures in finite samples.

correct specification of the model at significance level $\alpha \leq 0.215$ if the test statistic exceeds the critical value $cv_{n, \text{overid}} \equiv \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} \frac{1}{n} \text{trace} \left(\tilde{V} (I_p - WG(G'WG)^{-1}G') S (I_p - G(G'WG)^{-1}G'W) \right) \times (\Phi^{-1}(1 - \alpha/2))^2$, where we plug in sample analogues for all the unknown quantities.

¹⁴Note that in this case p continues to denote the total number of moments (“targeted” plus “non-targeted”), and in particular $\hat{G}$ should contain derivatives of both kinds of moments.

Model. We give a brief overview of the structural model here and refer the reader to [Alvarez and Lippi \(2014\)](#) for details. A firm sets prices on N products. The *desired* log prices for the products evolve in continuous time as N independent Brownian motions (without drift); however, the actual prices are fixed until the firm pays a fixed menu cost, at which point it may reset all N prices simultaneously. The firm’s profit depends negatively on the squared log deviation between the current and desired price, integrated over time and averaged across the N products. The model has $k = 3$ parameters: the number N of products, the volatility of the desired prices, and the scaled menu cost (relative to the curvature of the profit function).¹⁵ [Alvarez and Lippi \(2014\)](#) derive closed-form expressions for the frequency of price changes and the moments of the size of price changes. These are the moments we will match in the data.

Data. We empirically estimate the frequency and moments of price changes in scanner data from the supermarket chain Dominick’s.¹⁶ As described in detail in [Appendix B.1.4](#), we clean the data following [Alvarez, Le Bihan and Lippi \(2016\)](#), and in particular we focus on data from a single store. Unlike those authors, we exclusively use data on beer products, which arguably increases the interpretability of the results and makes the sample size more relevant for our subsequent simulation study. The final data set contains weekly prices on 499 beer products (Universal Product Codes, henceforth UPCs), observed for an average of 76 weeks per UPC. The total sample size is $n = 37,916$. When computing standard errors, we treat the price changes as i.i.d. across UPCs and time.

The $p = 4$ reduced-form moments that we match to the structural model are the average number of price changes per week as well as the empirical first, second, and fourth moments of the absolute log price changes (conditional on a nonzero change).¹⁷ We estimate the full-information covariance matrix of these moments using the usual nonparametric estimate (which depends on sample moments of price changes up to order 8). When applying our limited-information procedures, we use only the diagonal of this covariance matrix.

Results. We consider both just-identified and efficient specifications. We treat the number N of products as a parameter to be estimated, since there may not be a perfect correspondence between a UPC and the structural model’s notion of a “product”. The just-identified specification uses a weight matrix that attaches zero weight to the first moment of absolute price changes (i.e., the average), so that the three parameters are estimated from three moments (this estimator is available in closed form). We can then check whether

¹⁵In the notation of [Alvarez and Lippi \(2014\)](#), these parameters are n , σ , and $\sqrt{\psi/B}$, respectively.

¹⁶The data is provided by the James M. Kilts Center, University of Chicago Booth School of Business.

¹⁷Before computing moments, we subtract off the overall average log price change (conditional on a nonzero change), since the [Alvarez and Lippi \(2014\)](#) model abstracts from inflation.

PRICE SETTING APPLICATION: PARAMETER ESTIMATES

	Just-identified specification				Efficient specification		
	# prod.	Vol.	Menu cost	Over-ID	# prod.	Vol.	Menu cost
Full-info	3.012 (0.046)	0.090 (0.001)	0.291 (0.003)	0.002 (0.000)	3.255 (0.051)	0.089 (0.001)	0.305 (0.003)
Independ.	3.012 (0.167)	0.090 (0.001)	0.291 (0.010)	0.002 (0.001)	2.829 (0.091)	0.090 (0.000)	0.280 (0.006)
Worst case	3.012 (0.235)	0.090 (0.001)	0.291 (0.016)	0.002 (0.002)	2.786 (0.148)	0.090 (0.001)	0.278 (0.011)

Table 2.1: Estimates for the just-identified specification (uses only three moments for estimation) and the efficient specification (exploits all four moments for estimation). The rows correspond to full-information inference (exploits knowledge of $\hat{V}$), inference under independence (erroneously assumes that $\hat{V}$ is diagonal), and worst-case inference (exploits only diagonal of $\hat{V}$ without assuming off-diagonal elements are zero). Parameters: number of products (“# prod.”), volatility of desired log price (“Vol.”), scaled menu cost (“Menu cost”). Column “Over-ID” displays the error in fitting the non-targeted mean absolute price change moment, given the just-identified parameter estimates. Standard errors in parentheses.

the model provides a good fit for the “non-targeted” moment by carrying out the over-identification test proposed in [Section 2.4.2](#). The efficient specification exploits all four empirical moments, using either the conventional full-information one-step estimator or our limited-information one-step estimator in [Section 2.3.2](#). In addition to the full-information and limited-information procedures, we report results for a procedure that (erroneously) assumes that the four empirical moments are mutually independent.

[Table 2.1](#) shows that the limited-information standard errors are larger than the full-information ones, but they remain highly informative about the values of the structural parameters. In the efficient over-identified specification, the worst-case standard errors are at most 3.7 times larger than the corresponding full-information values. Importantly, all worst-case standard errors are arguably small relative to the economic magnitudes of the parameter estimates. Hence, taking a worst-case perspective still allows for informative inference. In this particular application, the standard errors that assume independence are mostly intermediate between the full-information and limited-information values.

Though limited-information inference is informative about the structural parameters themselves, there is a price to pay for the over-identification test in this application. In particular, [Table 2.1](#) shows that the limited-information test does not reject the validity of the non-targeted moment restriction, whereas the full-information test does reject (however, the economic magnitude of the moment violation is small, as the empirical moment equals 0.145 but the error in fitting the moment is only 0.002).¹⁸ This illustrates

¹⁸Consistent with the over-identification test, some of the full-information efficient parameter estimates are significantly different from the corresponding independence-based or limited-information estimates in [Table 2.1](#). If the model were correctly specified, there should be no statistically significant differences between estimates obtained with different weight matrices.

the principle that full-information inference is usually preferable if it is practically feasible. Of course, if we did not have access to the underlying supermarket scanner data, there would be no alternative to the limited-information analysis.

Simulation study. In [Appendix B.1.4](#) we show that our inference procedures perform well in a simulation study calibrated to the present empirical application. We simulate data from the [Alvarez and Lippi \(2014\)](#) model conditional on the estimated structural parameters. While our limited-information tests and confidence intervals have approximately correct size/coverage given the empirical sample size n (as do the full-information procedures), the procedures that erroneously assume independence between the reduced-form moments can over-reject/under-cover.

2.5.2 Heterogeneous agent New Keynesian model

Our second application estimates a heterogeneous agent New Keynesian general equilibrium macro model, following [McKay et al. \(2016\)](#) and [Auclert et al. \(2021\)](#). The matched moments are impulse response functions of macro time series and cross-sectional micro moments with respect to identified productivity and monetary policy shocks, as estimated by [Chang, Chen and Schorfheide \(2023\)](#) and [Miranda-Agrippino and Ricco \(2021\)](#). This is a medium-scale application with $k = 7$ parameters and $p = 23$ moments.

Though less efficient, impulse response matching estimation is more robust to modeling assumptions than full-information likelihood estimation.¹⁹ This is because—in the first-order approximation we consider—the impulse responses with respect to a monetary shock, say, do not depend on the exogenous processes for the other disturbances (e.g., shocks to the household discount rate).²⁰ Thus, our application only requires us to specify and estimate the exogenous processes for productivity and monetary disturbances. We remain agnostic about the number and nature of other shocks that may be driving the economy.

Likewise, our limited-information approach is simpler and less restrictive than other types of procedures that attempt to exploit more information. The only data inputs into our procedure are the impulse response point estimates and confidence intervals reported by [Chang et al. \(2023\)](#) and [Miranda-Agrippino and Ricco \(2021\)](#). We do not need access to the underlying data used in those papers, as would be required if one were to estimate the joint covariance matrix of all empirical moments via the bootstrap or GMM calculations.

However, the discrepancies between the different estimates are arguably small in economic terms.

¹⁹Likelihood procedures for estimation of heterogeneous agent models have been proposed by [Mongey and Williams \(2017\)](#), [Winberry \(2018\)](#), [Auclert et al. \(2021\)](#), and [Liu and Plagborg-Møller \(2023\)](#) among others.

²⁰Impulse responses to a monetary shock are computed by holding fixed all other exogenous shocks. As a result, the linearized impulse responses depend only on parameters of the monetary disturbance process as well as model parameters that govern the endogenous transmission mechanisms.

Unlike approaches based on bootstrapping or simulating data, we do not need to repeatedly re-run the impulse response estimation routines, and we do not need to fully model the relationship between the macro and micro data used by [Chang et al. \(2023\)](#) (e.g., by specifying all shocks).

Model. We employ the one-asset heterogeneous agent New Keynesian model described in [Auclert et al. \(2021, Appendix B.2\)](#); we refer to that paper for details, though we have made a couple of minor modifications to their model, as discussed below. Following [McKay et al. \(2016\)](#), the model features a continuum of heterogeneous households facing uninsurable idiosyncratic earnings risk. The households choose their work hours and amount of savings in a nominal Treasury bond. Monopolistically competitive firms set prices subject to a quadratic adjustment cost, yielding a New Keynesian Phillips curve. Households receive lump sum distributions of government interest revenue and firm profits. The central bank sets the nominal interest rate according to a Taylor rule that depends on inflation.²¹

The model is solved through a first-order linearization, using the numerical procedures developed by [Auclert et al. \(2021\)](#). That is, the moment function $h(\cdot)$ is computed numerically, and its derivatives are approximated via finite differences; no Monte Carlo simulation is involved. We study responses to (i) a productivity shock to the exogenous AR(2) process for the *log growth rate* of firms’ total factor productivity (TFP), and (ii) a monetary shock to an exogenous AR(2) process that enters as an additive disturbance in the Taylor rule.²²

Following [Auclert et al. \(2021\)](#), we limit ourselves to estimating structural parameters that do not affect the steady state of the model. This allows us to avoid repeatedly recomputing the steady state, though this would be feasible to do with moderate computational effort. The steady state parameters are fixed at the values assumed by [Auclert et al. \(2021, Table B.2\)](#). The $k = 7$ estimated parameters are: the Taylor rule coefficient on inflation, the slope of the Phillips curve, the three parameters in the AR(2) process for TFP growth, and the two autoregressive coefficients for the monetary disturbance.²³

Data. The empirical moments are obtained from two sets of Structural Vector Autoregression estimates of impulse responses to identified shocks.

Impulse responses with respect to TFP shocks are obtained from [Chang et al. \(2023, Fig. 9 and 11, blue](#)

²¹[Auclert et al. \(2021\)](#) allow the Taylor rule to depend also on output. We instead follow the Taylor rule used in those authors’ excellent GitHub repository: <https://github.com/shade-econ/sequence-jacobian>.

²²We thereby generalize the empirical specification of [Auclert et al. \(2021\)](#), which assumes an AR(1) monetary disturbance and no aggregate TFP disturbance. [Auclert et al.](#) additionally allow for shocks to government spending and markups; our impulse response matching estimation strategy is robust to the presence of such additional shocks, as noted above.

²³We do not need to estimate the standard deviation of the monetary shock, since this parameter does not affect the normalized impulse responses that we match (see below).

HETEROGENEOUS AGENT APPLICATION: PARAMETER ESTIMATES

Weight matrix	TR	PC	TFP			Monetary	
			AR1	AR2	Std	AR1	AR2
Diagonal	1.060 (0.100)	0.009 (0.008)	0.008 (0.146)	-0.040 (0.189)	0.006 (0.000)	0.702 (0.110)	0.075 (0.163)
Efficient	1.040 (0.074)	0.015 (0.006)	0.007 (0.137)	-0.017 (0.152)	0.006 (0.000)	0.697 (0.088)	0.014 (0.127)

Table 2.2: Structural parameter estimates with diagonal weight matrix (top row) and efficient weighting (bottom row). Parameters: Taylor rule coefficient on inflation (“TR”); slope of Phillips curve (“PC”); first and second autoregressive (“AR1” and “AR2”) and standard deviation (“Std”) parameters of TFP and monetary disturbance processes. Worst-case standard errors in parentheses.

lines). We use the responses of TFP itself and of GDP (output in the model), as well as the response of a cross-sectional moment estimated using data from the Current Population Survey (CPS): the fraction of people earning less than 2/3 of per capita GDP.²⁴ The sophisticated estimation method of [Chang et al. \(2023\)](#) takes into account statistical uncertainty arising from the limited sample sizes in the CPS. By relying directly on their reported results, our analysis inherits this desirable feature.

Impulse responses with respect to monetary shocks are obtained from [Miranda-Agrippino and Ricco \(2021, Fig. 3\)](#). We use the responses of industrial production (output in the model), the consumer price index (price level in the model), and the 1-year Treasury rate (annualized nominal interest rate in the model). Since our structural model is quarterly but the [Miranda-Agrippino and Ricco \(2021\)](#) data is monthly, we use the end-of-quarter impulse responses.

We focus on four impulse response horizons: the impact horizon, and the 1-, 2-, and 8-quarter horizons. When matching the model to the data, we take into account that the [Chang et al. \(2023\)](#) responses are with respect to a one-standard-deviation shock, while the [Miranda-Agrippino and Ricco \(2021\)](#) responses are normalized so that the Treasury rate increases by 100 basis points on impact.²⁵ Since both papers report Bayesian posterior quantiles, we appeal to the Bernstein-von Mises theorem and define the point estimates to be the reported posterior medians, while the standard errors are those implied by a normal approximation of the reported credible intervals.²⁶ In total, we have $p = 23$ empirical moments, as we discard the impact response of the bond rate, which is normalized to 1.

²⁴The factor 2/3 approximately adjusts for the average labor share, see [Chang et al. \(2023, Sec. 6.1\)](#).

²⁵[Chang et al. \(2023\)](#) actually consider a 3-standard-deviation shock, but we divide by 3.

²⁶Thus, if the length of the $1 - \alpha$ credible interval for θ_j is $\hat{L}_j$, we set $\hat{\sigma}_j = \hat{L}_j / (2\Phi^{-1}(1 - \alpha/2))$.

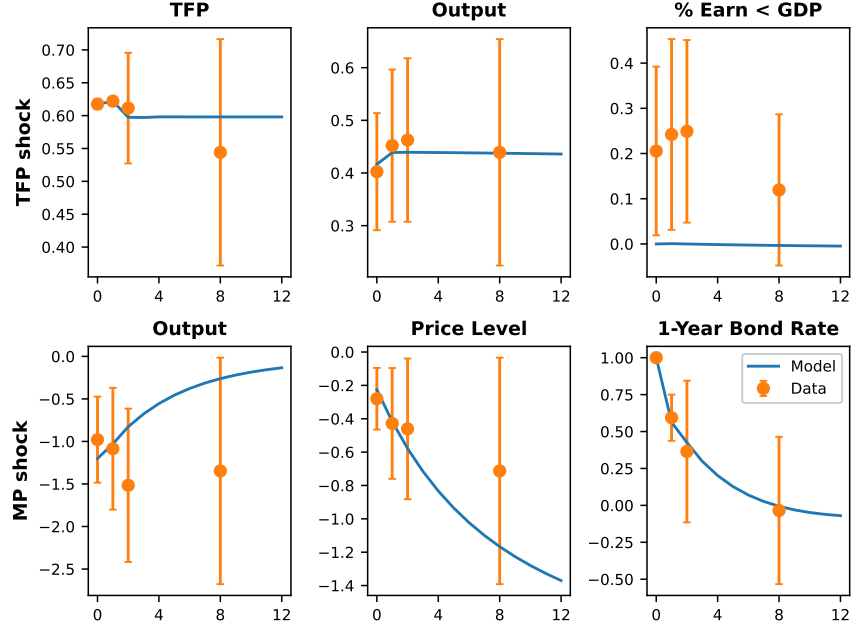


Figure 2.1: Model-implied impulse responses (thin curves) and corresponding empirical estimates (circles), with respect to a one-standard-deviation TFP shock (top row) or a monetary policy shock that raises the bond rate by 1 percentage point on impact (bottom row). Figure titles: response variables (“% Earn < GDP”: fraction of people earning less than 2/3 of per capita GDP). Vertical axis units: percentage points. Horizontal axis units: quarters. Vertical error bars: (shifted) 90% confidence intervals for the *differences* between the empirical and model-implied moments (not confidence intervals for the empirical moments themselves).

Results. The top row of Table 2.2 shows the parameter estimates obtained by using a diagonal weight matrix with $W_{jj} = 1/\hat{\sigma}_j^2$.²⁷ The Taylor rule coefficient on inflation is estimated to slightly exceed 1. The slope of the Phillips curve is positive but statistically insignificant at conventional significance levels. The TFP growth process is estimated to be close to white noise (i.e., the level of TFP is close to a random walk, as commonly assumed in the literature), while the monetary disturbance process has a half-life of about 2 quarters.

Figure 2.1 compares the model-implied and empirical impulse responses, at the parameter estimates discussed in the previous paragraph. We see that the model-implied impulse responses of output to a monetary shock are too small in magnitude relative to the data at the 2- and 8-quarter horizons, while the opposite is true for the responses of the price level with respect a monetary shock. Moreover, the model fails to generate nontrivial dynamics in the fraction of low-wage earners in response to a TFP shock. To test

²⁷We run gradient-based numerical optimizations from 100 different starting values and report the overall optimum. One vector of starting values equals the posterior mode estimates $\hat{\theta}_i^{\text{ABRS}}$ from Auclert et al. (2021, Table F.III), and imposing a simple random walk for TFP. The 99 other starting values are simulated uniformly at random from intervals of the form $[0, 2\hat{\theta}_i^{\text{ABRS}}]$, or $[-0.99, 0.99]$ in the case of AR1 coefficients.

whether these disparities are too large to be explained by statistical noise, we conduct the over-identification test proposed in [Section 2.4.2](#). The vertical error bars in [Figure 2.1](#) show the 90% confidence intervals for the *differences* between model-implied and empirical moments, centered at the empirical moments for visual convenience. Three of the model-implied impulse responses for the fraction of low-wage earners fall outside their respective intervals, though only marginally so. While this points to misspecification of either the structural model or the reduced-form VAR models used to generate the empirical impulse responses, we note that the *joint* test of the validity of all $p = 23$ moments does not reject at the 10% level.²⁸

The efficient (one-step) parameter estimates in the bottom row of [Table 2.2](#) demonstrate the benefit of optimally weighting the moments as described in [Section 2.3.2](#).²⁹ In particular, the t-statistic for the slope of the New Keynesian Phillips curve increases to 2.32, from 1.10 previously. Thus, our limited-information approach yields informative inference about a parameter that is often viewed as difficult to pin down in the data ([Mavroeidis, Plagborg-Møller and Stock, 2014](#)). More generally, the efficient standard errors in the bottom row of [Table 2.2](#) are 6–26% smaller than the non-efficient ones in the top row.

2.6 Conclusion

We computed simple, sharp, and informative upper bounds on the standard errors of structural parameter estimates when the correlation structure of the matched empirical moments is not fully known. In addition, we proposed an efficient moment weighting procedure in the over-identified case, as well as valid tests of parameter restrictions and over-identifying restrictions. The required inputs are minimal: Other than being able to evaluate the mapping from structural parameters to model-implied moments (at least numerically), we just need the empirical moment estimates and their individual standard errors. Our procedures are computationally tractable even in settings with many moments and/or parameters. A code suite is available online (see [Footnote 1](#)).

We believe our limited-information approach is useful for applied researchers who match their models to moments obtained from several different data sources, estimation methods, or previous papers. Our methods obviate the need to estimate the correlation structure across the various moments, which is sometimes difficult or impossible. Even when the moment correlation structure is in principle estimable, our methods may be helpful, since marginal standard errors for individual moments are typically much easier to obtain from standard econometric software than it is to figure out the joint distribution of all moments, as illustrated in

²⁸The test statistic in [Footnote 13](#) equals 24.58 with critical value 56.99.

²⁹[Table B.3 in Appendix B.1.5](#) shows which moments are optimally selected to estimate the various parameters.

our empirical applications. Moreover, the limited-information procedures can be used to gauge whether it is worthwhile to expend the additional effort required for full-information analysis.

Chapter 3

Tilting VAR Forecasts on Marginal Information

3.1 Introduction

Given a VAR with a small number of aggregates, one might wish to identify accuracy gains from exploiting external information not in the model and, ultimately, adjust forecasts on the basis of that information if determined useful. That information could include survey data (Tallman and Zaman, 2020) or leading indicators for the VAR’s outcomes (Krüger, Clark and Ravazzolo, 2017). Operationally, one could either expand and re-estimate the model or—holding fixed the original model—use entropic tilting as a post-processing step to recenter the forecast distribution based on the additional information. While clearly related, the precise relationship is not *ex ante* obvious, and there are tradeoffs to each approach. On the one hand, entropic tilting is a computationally cheap post-estimation step, whereas model reformulation and re-estimation is more analytically and computationally involved. Moreover, the post-estimation nature of tilting also allows clean separation of the core VAR dynamics—which one typically wishes to carefully model, estimate using a long sample, and use for dynamic structural analysis—from purely predictive covariates, which may only be of interest to the extent that they improve accuracy. On the other other hand, tilting can be entirely *ad hoc*, while re-estimation (under proper penalization) is the data-driven “ideal” from an econometric point of view.

To relate both re-estimation and tilting, I outline a particular implementation of entropic tilting for

VARs that exploits the additional covariates, and then I prove its equivalence to re-estimation with those covariates. The implication is that tilting need not be completely ad hoc. But rather, one can design tilting procedures that exactly match or, at minimum, mimic the ideal of re-estimation. Moreover, the procedure itself consists of a sequence of intuitive steps that extract and highlight the marginal information introduced by the additional information. Rather than throwing away the forecast distribution of the original core VAR and replacing with an ad hoc choice, the forecast is refined based on cleanly extracted marginal information.

The procedure itself consists of a sequence of simple intuitive steps, which follow from a Bayesian version of the Frisch-Waugh-Lovell theorem and omitted variable bias formulas. Specifically, given the additional predictive regressors, the recommendation of this paper is to (1) project out the regressors already included in the core BVAR (i.e. residualize the additional covariates) and then (2) regress the core VAR's forecast errors on the result of step 1. This pins down how one should adjust the forecast distribution of the core VAR based on marginal external information. I use that relationship to adjust the forecast distribution of the core VAR in a final tilting step. Lastly, in the initial two steps, I emphasize how one can and should use dummy observations, which introduce proper penalization and ultimately ensure that the adjusted forecast distribution exactly matches (or mimics) that under re-estimation.¹ As a consequence, this procedure is data-driven in its adjustment of the forecast distribution, but fully implemented as a post-estimation step via tilting.

After proving the analytical equivalence results which justify and guide the recommended procedures, I illustrate the procedure in an out-of-sample forecasting exercise. The exercise borrows from the nowcasting application of Krüger et al. (2017). I start with a core BVAR for quarterly GDP growth, inflation, unemployment, and the T-Bill rate. I then use additional higher frequency data as additional regressors to refine and tilt the forecast of the original core BVAR. Whereas Krüger et al. (2017) and other authors develop auxiliary nowcasting models for each outcome in the core BVAR, I use my recommended procedure in a single step. In all cases, the adjusted forecast distributions from variations of the procedure are more accurate than the baseline forecast distribution from the core BVAR. Further, for all but the T-Bill (which is somewhat unique relative to the other variables, as I will discuss), the procedure performs comparably to the nowcasting approach of Krüger et al. (2017). This suggests that one need not develop auxiliary nowcast models outcome-by-outcome, but rather can employ a single postestimation step to refine the forecast distribution based on the additional data. This also has the advantage of exploiting a wider range of data for forecast refinement.

¹By “re-estimation”, I am implicitly referring to re-estimation of an encompassing model that includes all covariates.

Literature The tilting literature in macroeconomics began with [Robertson, Tallman and Whiteman \(2005\)](#) and [Cogley, Morozov and Sargent \(2005\)](#), which introduced the procedure for VAR forecast distributions. One strand of the literature has used tilting to conduct economically-motivated scenario analyses as in [Crump, Eusepi, Giannone, Qian and Sbordone \(2021\)](#) and [Antolín-Díaz, Petrella and Rubio-Ramírez \(2021\)](#), or to impose economic constraints as in [Giacomini and Ragusa \(2014\)](#). The latter paper is notable because, like this paper, it also establishes an analytical result to justify a specific tilting procedure. However, the justification of [Giacomini and Ragusa \(2014\)](#) rests on economic theory. They show that if economic constraints are true in population, imposing them via tilting can be expected to improve accuracy. In contrast, I justify tilting my tilting procedure on the basis of its relation to re-estimation, rather than economic theory.

Another strand of the tilting literature (which includes this paper) has explored the forecast accuracy improvements that come from tilting based on additional information. Different studies have varied the input data used in tilting: [Bańbura, Brenna, Paredes and Ravazzolo \(2021\)](#) consider survey data, [Krüger et al. \(2017\)](#) and this paper consider higher frequency macroeconomic data or nowcasts, while [Tallman and Zaman \(2020\)](#) consider both. Other examples include [Doh \(2017\)](#), [Koop, McIntyre and Mitchell \(2020\)](#), [Galvão, Garratt and Mitchell \(2021\)](#), and [Knotek and Zaman \(2019\)](#), and the many references in these studies.

My paper differs in that I explicitly link tilting and re-estimation, and in doing so, recommend an alternative tilting implementation. The existing empirical literature treats tilting as ad hoc, separate from estimation, and *fully* tilt the forecast distribution of outcomes towards values suggested by auxiliary nowcast models or surveys. In other words, they amount to a specification change at near-term horizons (in the case of nowcasting) or replacement of estimation with judgment. In contrast, my recommended procedure is equivalent or similar to estimation of an expanded model, and explicitly characterizes the marginal information in the additional covariates. This has the advantage of making clear the source of forecast adjustment and improvement.

My main analytical result is most closely related to [Osiewalski \(2011\)](#), which offers a quite complete treatment of the Frisch-Waugh-Lovell Theorem for parameter estimation in a Bayesian context. In this paper, I instead pay extra attention to the forecast distribution, as my main analytical result establishes equivalence of the tilted forecast distribution and re-estimation forecast distribution. I also focus on operationalizing the FWL Theorem for the purpose of forecast adjustment in a macroeconomic time series forecasting context. This involves the use of dummy observations to encode the priors, which have a long tradition in macroeconomic forecasting dating back to [Litterman \(1980\)](#). While these dummy variables equivalently implement

the same sorts of priors considered in Osiewalski (2011), they allow me to define a forecast adjustment procedure that can *reuse* the original estimation of the short regression, without recomputing or resampling the posterior. Finally, my adjustment procedure also requires stating a Bayesian equivalent of omitted variable bias, which follows from the same logic and setup of Osiewalski (2011), and which I make explicit.

Finally, my paper is related to growing literature that uses machine learning to improve forecasts, for two reasons.

First, arbitrary machine learning models can potentially be used to detect and exploit the predictability provided by additional regressors. In particular, my procedure adjusts the forecast distribution based on the estimated relationship between the core VAR’s forecast errors and the marginal information contained in additional predictors. Intuitively, if the forecast errors can be predicted by the marginal information, one can exploit that relationship to improve accuracy. And in principle, *any* predictive model can be used in this step, including machine learning models. While this typically breaks exact equivalence as I discuss, my equivalence result still serves as a rough yet practical justification and provides guidance on how to correctly account for the penalization imposed in the original VAR. See Masini, Medeiros and Mendes (2023) for a survey of the use of machine learning for macroeconomic forecasting. Specific recent examples also include Marcelo C. Medeiros and Zilberman (2021), Yoon (2021), and Lindenmeyer and da Silva Torrent (2023). While these papers dispense with a core BVAR altogether, their insights can help guide how the forecast can be adjusted using the residuals in the time series setting.²

Second, the use of forecast errors for model improvements is related to the idea of boosting in machine learning.³ Luo and Spindler (2017) discuss boosting algorithms and asymptotic results in a regression context, with primary application in the causal inference literature, e.g. Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2018). As the authors emphasize, boosting uses regression residuals and an iterative process to select predictors to use in a final (cross-sectional) predictive model, iterated until some stopping criterion is met. Therefore, these boosting schemes are variable selection schemes that impose sparsity. In this paper, I focus on the macroeconomic time series setting, and my recommended procedure instead uses residuals to drive an adjustment under *shrinkage*, which matches the common Minnesota prior more commonly used and appropriate in macroeconomic forecasting (Giannone, Lenza and Primiceri, 2021). My approach also makes explicit and formal connections to finite-sample re-estimation under a Bayesian prior, whereas boosting tends to rely on ad hoc stopping criteria, justified by asymptotics.

²I retain the BVAR as the fixed, core model because I have in mind applications where researchers might like to use the BVAR for standard dynamic causal analysis, but seek to refine its forecasts.

³I thank Mikkel Plagborg-Møller for pointing out this connection.

3.2 Setup

In this section, I provide the setup necessary to define and prove the main results. [Section 3.2.1](#) defines the core VARX estimated using only a subset of available information, as well as an implicit extended VARX that uses additional covariates. Though the latter typically will not be estimated, it is necessary to explicitly define this full-information VARX so I can define its posterior predictive (forecast) distribution, which my procedure will match (in mean) in the leading case. Comparing posterior distribution requires taking a stand on priors. [Section 3.2.2](#) defines the prior families this paper will cover. [Section 3.2.3](#) gives a brief overview of entropic tilting, which is the method my procedure uses to adjust the forecast distribution of the core VARX, once the information is properly extracted from additional covariates.

3.2.1 Core and Full-Information VARX Models

This section establishes enough notation to characterize the partial-information core VARX (whose forecast will be adjusted), the implicit full-information VARX, as well as the posterior and posterior predictive (forecast) distributions of both. Therefore, this section names all of the main analytical objects of interest and will allow me to clearly distinguish between the distinct posteriors and forecast distributions of the core and explicit VARXs.

Definition 18. (Core/“Short” VARX) A VARX with n_1 outcomes collected in vector y_t , p lags, and m_1 additional regressors z_t , can be written

$$y_t = A_1 y_{t-1} + \cdots + A_p y_{t-p} + C z_t + u_t \quad \text{where} \quad u_t | y_{t-1}, \dots, y_{t-p}, z_t \sim \mathcal{N}(0, \Sigma_A)$$

This VARX can be stacked into a standard linear model as follows

$$\begin{pmatrix} y'_1 \\ \vdots \\ y'_T \end{pmatrix} = \begin{pmatrix} y'_0 & \cdots & y'_{1-p} & z'_1 \\ \vdots & \ddots & \vdots & \vdots \\ y'_{T-1} & \cdots & y'_{T-p} & z'_T \end{pmatrix} \begin{pmatrix} A'_1 \\ \vdots \\ A'_p \\ C' \end{pmatrix} + \begin{pmatrix} u'_1 \\ \vdots \\ u'_T \end{pmatrix}$$

$$\underset{(T \times n_1)}{Y} = \underset{(T \times K_1)}{X_1} \underset{(K_1 \times n_1)}{A} + u \tag{3.1}$$

where $K_1 = n_1 p + m_1$, Note that this might simply be a VAR(p) where C consists of only a constant term,

and no additional regressors. If there are additional regressors in z_t , they could be deterministic or exogenous regressors.

Definition 19. (Prior, Posterior, and Forecast Distribution for “Short” Regression) The model in [Definition 18](#) is characterized by parameters (A, Σ_A) and has an associated likelihood $\ell_A(Y|A, \Sigma_A)$. See remark below on the treatment of initial conditions. Estimation of these parameters is commonly done using Bayesian methods, which entails specifying a prior $\pi_A(A, \Sigma_A)$ and deducing the posterior

$$p_A(A, \Sigma_A|Y) \propto \ell_A(Y|A, \Sigma_A) \cdot \pi_A(A, \Sigma_A)$$

where the “ A ” subscript is used to distinguish this core VARX posterior from the full-information posterior I define and consider below. Throughout this paper, I suppress conditioning on covariates in the expression posterior to save space and avoid confusion.

Given a set of parameters, if the additional regressors in z_t are deterministic, one can define the forecast distribution for h periods ahead

$$f_A(y_{T+1:T+h}|A, \Sigma_A, Y)$$

as well as the posterior predictive (forecast) distribution

$$f_A(y_{T+1:T+h}|Y) = \int f_A(y_{T+1:T+h}|A, \Sigma_A, Y) p_A(A, \Sigma_A|Y) dA d\Sigma_A.$$

If the VARX includes exogenous regressors so that iterated multi-step ahead forecasts are not available, one can define the above for $h = 1$ and resort to direct forecasting for longer horizons (see [Section 3.3.3](#) for discussion).

Remark. (Initial Conditions) Though I do not distinguish in notation, I condition on initial conditions $(y_0, \dots, y_{1-p})$ throughout this paper. As a consequence, all posterior inference and posterior predictive distributions are also conditional on initial conditions. This is typically done when VAR outcomes are nonstationary, e.g. for the level of GDP, the level of the price index. It is also commonly done in practice, even when outcomes are stationary. However, in a pure VAR setting, when the variables are known to be stationary, it is straightforward to derive the unconditional distribution of initial conditions as function of the parameters and include this piece in the likelihood and posterior. See [Chib and Greenberg \(1994\)](#), who discuss this for the AR model, which generalizes to VAR models.

I do not pursue this, first, because estimation of BVARs in levels or with near nonstationary outcomes is quite common (Giannone, Lenza and Primiceri, 2015). Second, as discussed in Chib and Greenberg (1994), incorporating initial conditions leads to a nonstandard posterior distribution, and the equivalence results below rely on analytical characterization of the posterior distribution.

Definition 20. (Full-Information/Extended/“Long” VARX) If another block of regressors are available, $X_2 \in \mathbb{R}^{T \times K_2}$, one could define an extended “full information” VARX

$$Y = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \underbrace{\begin{pmatrix} B_1 \\ B_2 \end{pmatrix}}_B + \varepsilon \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma_B) \quad (3.2)$$

These additional regressors are unrestricted and could consist of deterministic regressors, exogenous regressors not included in the original VARX, or multiple lags of other time series that could conceivably have been modeled within original VARX.

Comparing to Definition 18 and borrowing terminology from results on omitted variable bias, Expression 3.2 is a “long” regression counterpart to the “short” regression in Expression 3.1.

Remark. Again, this expanded VAR is introduced mainly as a useful artificial construct. I do not expect the econometrician to estimate this model. Rather, I define it so I can show that the adjusted forecast of my recommended procedure, which avoids re-estimation, corresponds to the forecast distribution of a well-defined model.

For that reason, I also ignore complications of different sample sizes between (Y, X_1) and X_2 . That is an issue for practical implementation that can be handled by imposing more or less shrinkage on the additional regressors to account for, respectively, a smaller or larger relative sample size. Moreover, the key ideas are easiest to communicate ignoring that possibility.

Definition 21. (Prior, Posterior, and Forecast Distribution for “Long” Regression) The model in Definition 20 is characterized by parameters (B, Σ_B) and has an associated likelihood $\ell_B(Y|B, \Sigma_B)$, where the “B” subscript is used as shorthand to signify the “long” regression. As in Definition 19, when estimating the expanded/long regression using Bayesian methods, one defines a prior $\pi_B(B, \Sigma_B)$ and computes the posterior $p_B(B, \Sigma_B|Y)$

$$p_B(B, \Sigma_B|Y) \propto \ell_B(Y|B, \Sigma_B) \cdot \pi_B(B, \Sigma_B)$$

There is also a corresponding forecast distribution for one period ahead given a set of parameters, $f_B(y_{T+1}|B, \Sigma_B, Y)$, as well as the forecast distribution that averages over the posterior distribution, $f_B(y_{T+1}|Y)$.

Since X_2 will contain exogenous (rather than deterministic) regressors, I only consider the one step ahead forecast distribution rather than an iterated multi-step, which is not available without another forecast model for exogenous regressors. In [Section 3.3.3](#), I discuss how adjustment of one step ahead forecasts in my procedure will nonetheless propagate to longer horizons, as well as suggest further modifications based on direct forecast that one could do if concerned that the standard propagation is insufficient.

3.2.2 Prior Specifications

While the previous section offered notation for priors and posteriors in general, I use this section to make explicit the functional form of priors that I will consider in this paper. This is required because the leading exact equivalence result of this paper holds under one particular (common and standard) prior family, which I introduce first. I then conclude the section by introducing a (common and standard) generalization of that prior family. Though an exact equivalence result will not hold for this family, it can still be used while retaining much of the intuition of exactly equivalent prior family. Therefore, I define it for later analysis and discussion, along the same lines as the first prior family.

Definition 22. (Normal-IW Kronecker Conjugate Prior for Core VARX) A common prior specification in the BVAR literature is the conjugate (dependent) Normal-IW prior

$$\pi_A(A, \Sigma_A) = \pi_A(A|\Sigma_A)\pi_A(\Sigma_A)$$

$$\pi_A(A|\Sigma_A) = \pi_A(\text{vec}(A)|\Sigma_A) \sim \mathcal{N}(\text{vec}(A_0), \Sigma_A \otimes \Phi_{0,A})$$

$$\pi_A(\Sigma_A) \sim IW(S_0, \alpha_0)$$

given econometrician-specified tuning parameters $(A_0, \Phi_{0,A}, S_0, \alpha_0)$, where $IW(S, \alpha)$ is an inverse-Wishart distribution, with functional form defined in [Appendix C.2.4](#).

This specific choice of prior implies an associated posterior distribution $p_A(A, \Sigma_A|Y)$ of the same conditional form. See Proposition 25 in [Appendix C.2.2](#) for the exact functional forms, which are standard results from the literature ([Karlsson, 2013](#)).

Definition 23. (Normal-IW Kronecker Conjugate Prior for Extended VARX) For the extended VARX,

one can define a corresponding conjugate (dependent) Normal-IW prior

$$\begin{aligned}\pi_B(B, \Sigma_B) &= \pi_B(B|\Sigma_B) \cdot \pi_B(\Sigma_B) \\ \pi_B(B|\Sigma_B) &= \pi_B(\text{vec}(B)|\Sigma_B) \sim \mathcal{N}\left(\underbrace{\text{vec}\begin{pmatrix} B_{0,1} \\ B_{0,2} \end{pmatrix}}_{B_0}, \Sigma_B \otimes \underbrace{\begin{pmatrix} \Phi_{0,11} & \Phi_{0,12} \\ \Phi_{0,21} & \Phi_{0,22} \end{pmatrix}}_{\Phi_0}\right) \\ \pi_B(\Sigma_B) &\sim IW(S_0, \alpha_0)\end{aligned}$$

given econometrician-specified tuning parameters $(B_0, \Phi_0, S_0, \alpha_0)$.

In defining this prior, it is natural to think of this as an extension of the prior in [Definition 22](#), augmented to handle the additional covariates. In particular, if one takes $B_{0,1} = A_0$ and $\Phi_{0,11} = \Phi_{0,A}$, as well as identical (S_0, α_0) , the prior for the first block of coefficients is identical. Then $B_{0,2}$ and $\Phi_{0,22}$ characterize the prior distribution for the added second block of coefficients/covariates, while $\Phi_{0,21}$ captures prior correlation between the coefficients.⁴

Again, this specific choice of prior implies an associated posterior distribution $p_B(B, \Sigma_B|Y)$ with the same conditional form as above.

In the main main equivalence result of [Section 3.3.1](#), I focus on priors with Kronecker structure because they offer a sharp characterization of equivalence between tilting and re-estimation. And even though restrictive, they provide strong intuition and a solid foundation, from which one can consider reasonable departures from exact equivalence.

For one of those departures, I will use a common, alternative prior family from the literature that lacks the Kronecker form. This family is useful because the Kronecker structure in [Definitions 22](#) does impose a non-trivial restriction on the core VARX.⁵ Therefore, this alternative prior form can be preferable empirically. To ultimately test this, I now define the prior, later develop an alternative forecast adjustment procedure under this prior in [Section 3.3.2](#), and compare its performance in the empirical application of [Section 3.4](#).

Definition 24. (Normal-IW Non-Kronecker Conjugate Priors for Core and Expanded VARX) Suppose that

⁴ $\Phi_{0,21}$ is often taken to be zero in practice, but I allow nonzero for full generality.

⁵Namely, the prior variance-covariance matrices for coefficients in the different equations for the different outcomes are *proportional*, and so one cannot impose, for example, more shrinkage on cross-lags than own-lags for a given equation because this would violate proportionality across equations. See also [Karlsson \(2013\)](#) for an extensive discussion and reference on more alternatives to the prior with Kronecker structure.

the core VARX is estimated with the following prior

$$\pi_A(A, \Sigma_A) = \pi_A(A)\pi_A(\Sigma_A)$$

$$\pi_A(A) = \pi_A(\text{vec}(A)) \sim \mathcal{N}(\text{vec}(A_0), \Omega_{0,A})$$

$$\pi_A(\Sigma_A) \sim IW(S_0, \alpha_0)$$

The difference from [Definition 22](#) is that the prior over coefficients A is now independent of Σ_A . The variance no longer has Kronecker structure that scales with Σ_A .

For the implicit full-information VARX that also includes X_2 , suppose

$$\pi_B(B, \Sigma_B) = \pi_B(B)\pi_B(\Sigma_B)$$

$$\pi_B(B) = \pi_B(\text{vec}(B)) \sim \mathcal{N}\left(\underbrace{\text{vec}\begin{pmatrix} B_{0,1} \\ B_{0,2} \end{pmatrix}}_{B_0}, \underbrace{\begin{pmatrix} \Omega_{0,11} & \Omega'_{0,21} \\ \Omega_{0,21} & \Omega_{0,22} \end{pmatrix}}_{\Omega_0}\right)$$

$$\pi_B(\Sigma_B) \sim IW(S_0, \alpha_0)$$

This can be viewed as an extension of the prior for the core VARX.

3.2.3 Entropic Tilting

In this section, I define entropic tilting since it is central to the recommended procedure. Given an initial probability distribution over outcomes, entropic tilting replaces that distribution with a modified distribution that satisfies a set of econometrician-supplied constraints.

In this paper, the initial distribution is the multi-step ahead posterior predictive (joint forecast) distribution, $f_A(y_{T+1:T+h}|Y)$ from the core VARX. My recommended procedure will generate constraints for modifying this distribution based on the information contained in the additional covariates X_2 that were not included in the original core VARX.

Definition 25. (Entropic Tilting) The method of entropic tilting solves for a new distribution $f_*(y_{T+1:T+h})$ given an initial distribution $f(y_{T+1:T+h})$ and a set of constraints $\mathbb{E}_{f_*}[g(y_{T+1:T+h})] = \bar{g}$ (which can be a vector

of constraints) that must be satisfied under the new distribution f_* , on average. The econometrician must fully specify $g(\cdot)$ and $\bar{g}$.

Because there are typically many candidate distributions that would satisfy the constraint, a criterion is needed to select among the corresponding family of compatible distributions. Entropic tilting is the name given to the procedure that uses the Kullback-Leibler divergence because of the connection between that divergence metric and relative entropy. More formally,

$$f_* = \operatorname{argmin}_{\tilde{f}} \{KLIC(\tilde{f}, f) \mid \mathbb{E}_{\tilde{f}}[g(y_{T+1:T+h})] = \bar{g}\} \quad (3.3)$$

where $KLIC(\tilde{f}, f)$ is the Kullback-Leibler divergence metric. The resulting “tilted” distribution f_* depends upon the starting distribution f and the constraint(s), and it represents the closest distribution to the starting distribution f (in a KL divergence sense) while still satisfying the constraints, provided a solution exists.⁶

Remark. (Implementation) There are two ways to implement entropic tilting.

The most common and most relevant for this paper follows [Robertson et al. \(2005\)](#) and reweights draws from the initial forecast distribution $f(y_{T+1:T+h})$ so that, under reweighting, the draws satisfy the constraint. Formally, suppose one has draws $\{y_{T+1:T+h}^{(d)}\}_{d=1}^D$ from the baseline forecast distribution. Note that each draw $y_{T+1:T+h}^{(d)}$ is an h -period path, and the set of draws implicitly have empirical probability weights $p_d = 1/D$. To turn these into draws from $f_*(y_{T+1:T+h})$, replace the empirical weights with computed weights $p_d^* \in (0, 1)$ such that

$$\sum_{d=1}^D p_d^* g(y_{T+1:T+h}^{(d)}) = \bar{g}$$

[Robertson et al. \(2005\)](#) describes how to compute these weights $\{p_d^*\}_{d=1}^D$ by solving a discrete-distribution counterpart of Problem 3.3. These weights are used whenever computing moments with the draws $\{y_{T+1:T+h}^{(d)}\}_{d=1}^D$, which is equivalent to computing moments with genuine draws from $f_*(y_{T+1:T+h})$.

Alternatively, it is sometimes possible to solve the tilting problem in Expression 3.3 analytically for f_* and, given the analytical characterization, produce new draws from $f_*(y_{T+1:T+h})$ directly.⁷ This may be

⁶See [Tallman and West \(2022\)](#) for a review of tilting and conditions for existence of a solution.

⁷[Antolín-Díaz et al. \(2021\)](#) is a relevant example from the macroeconometric literature where they show that if the initial distribution is normal and the constraints are a linear function of $y_{T+1:T+h}$, the resulting tilted distribution is normal. Draws can be generated from the new distribution via direct sampling or the simulation smoother (see Online Appendix E), which may be more efficient. [Tallman and West \(2022\)](#) review results for a wider class of starting distributions, namely exponential family distributions.

preferable when the constraints in Expression 3.3 are satisfied only in the tails of the initial distribution f , implying extreme weights $\{p_d^*\}$ on a small number of draws. In that case, directly drawing from the target distribution avoids the problem of extreme reweighting of initial draws (even if calling into question the prudence of these adjustments).

In this paper, I will consider only the first approach for three reasons. First, one important focus is on reuse of artifacts from initial estimation of the core VARX. The first procedure involves reweighting draws $f_A(y_{T+1:T+h}|Y)$, which one likely already has from constructing the initial baseline forecast distribution. Second, the first approach allows the simplest formulation of the procedure. While it is possible to modify the procedure to fit the second approach, those modifications distract from the main ideas. Finally, if the uncertainty bands of the original core VARX are well-calibrated in terms of coverage, the additional covariates are unlikely to lead to extreme weights on a few draws in tails of the distribution.

3.3 Tilting Procedures for Forecast Adjustment

This section recommends procedures for adjusting the forecast distributions of the core VARX, given additional covariates. As a starting point, Section 3.3.1 defines a baseline procedure and proves an equivalence result, which serves as a justification for all procedures I recommend. I state precisely the assumptions on the prior required for this equivalence to hold. From that starting point, I then define two variations on the procedure in Section 3.3.2. These variations give examples for how one can extend the procedure to potentially improve forecast accuracy as well as cover cases where the original assumptions on the prior do not hold exactly. In either case, both build on the intuition and spirit of the baseline case, and I will test their performance in Section 3.4. Lastly, Section 3.3.3 makes explicit the implications for longer horizon forecasting of the adjustment procedures.

3.3.1 Tilting and Re-Estimation Equivalence

Throughout this section, I continue with the setup of Section 3.2.1, which defined the core VARX and the implicit full-information VARX given additional covariates. Because the main equivalence result of this section depends upon the prior, I first state the required assumptions on the prior for re-estimation equivalence, recalling the prior definitions of Section 3.2.2.

Assumption 12. (Priors for Core and Expanded VARX) Suppose that one has estimated the core VARX using the prior of Definition 22, given econometrician-specified values for the tuning parameters $(A_0, \Phi_{0,A}, S_0, \alpha_0)$.

Further suppose that one has access to additional covariates X_2 . These can be used to define an extended VARX that employs the prior of [Definition 23](#). Assume $B_{0,1} = A_0$ and $\Phi_{0,11} = \Phi_{0,A}$, as well as identical S_0 and α_0 . The values for the remaining tuning parameters ($B_{0,2}, \Phi_{0,22}, \Phi_{0,21}$) can be chosen freely by the econometrician, as they would choose them if re-estimating the expanded VARX, subject to ensuring Φ_0 is symmetric positive definite.

Remark. (Justifying the Assumption on S_0 and α_0) In principle, adding covariate X_2 will reduce the forecast error variance, i.e. the difference $\Sigma_A - \Sigma_B$ is likely a symmetric positive definite matrix. Therefore, maintaining the same S_0 and α_0 between priors for the core VARX and extended VARX seems inappropriate.

However, I note that S_0 and α_0 are often chosen in practice without reference to the covariates and chosen to implement a diffuse prior on the forecast error variance. See [Giannone et al. \(2015\)](#), for example. Forecasters often have stronger beliefs about the dynamics encoded by A and B , rather than strong prior beliefs about the forecast error variance. Moreover, unlike the coefficient matrices, Σ_A and Σ_B do not scale in dimension with additional covariates, only with the number of outcomes N . Therefore, there is not a strict requirement to adjust between the core and full-information VARXs.

Proposition 15. (Posterior and Posterior Predictive Distributions) *Under [Assumption 12](#), the means of the marginal posterior distributions of the coefficients for the core and expanded VARXs satisfy*

$$\begin{aligned}\mathbb{E}_A[A|Y] &= \bar{A} = [\Phi_{0,A}^{-1} + X_1'X_1]^{-1}[\Phi_{0,A}^{-1}A_0 + X_1'Y] \\ \mathbb{E}_B[B|Y] &= \bar{B} = [\Phi_{0,B}^{-1} + X'X]^{-1}[\Phi_0^{-1}B_0 + X'Y]\end{aligned}$$

where $\mathbb{E}_A[\cdot | Y]$ is the expectation taken with respect to the posterior distribution $p(A, \Sigma_A | Y)$; similar for $\mathbb{E}_B[\cdot | Y]$. As a consequence, the mean of the posterior predictive distributions for the one-step ahead forecast distributions are

$$\mathbb{E}_A[y_{T+1}|Y] = x_{T+1,1}\bar{A} \tag{3.4}$$

$$\mathbb{E}_B[y_{T+1}|Y] = x_{T+1,1}\bar{B}_1 + x_{T+1,2}\bar{B}_2 \tag{3.5}$$

Proof. See Proposition 25 in [Appendix C.2.2](#). □

Remark. These marginal posterior means can be computed analytically according to the formulas above or via OLS with dummy observations, as outlined in Corollary 4 in [Appendix C.2.1](#). They could also be computed up to Monte Carlo error by averaging across draws from posterior sampling, as outlined in

Proposition 25 of Appendix C.2.2. For all analytical results of this section, I ignore Monte Carlo error for simplicity and because it can be made arbitrarily small with enough draws.

Proposition 16. (Relating Forecast Residuals) *Let $\hat{u} = Y - X_1\bar{A}$ denote the vector of in-sample forecast residuals from the core VARX and $\hat{\varepsilon} = Y - X_1\bar{B}_1 - X_2\bar{B}_2$ denote the forecast residuals from re-estimation of the implicit expanded VARX. Then*

$$\hat{u} = \hat{\varepsilon} + \underbrace{(X_2 - X_1P)}_{\hat{v}} \bar{B}_2 - X_1R,$$

where $\hat{v}$ are projection residuals capturing the marginal information of X_2 (extracted with shrinkage adjustment), P is a projection coefficient matrix, and R is an adjustment term for prior correlation in coefficients across blocks (and which equals zero under prior independence).⁸

Proof. See Theorem 4 in Appendix C.3.2. □

In Proposition 15, the mean and distributions between models clearly differ, and in Proposition 16, the short-regression residuals are essentially equal to the long-regression residuals plus additional noise, due to leaving out marginal information contained in X_2 , as for OLS. The motivating idea of this section and paper is that one could use entropic tilting to recenter the posterior predictive distribution to satisfy Expression 3.5, rather than Expression 3.4. In other words, one could use Expression 3.5 as the constraint and update the forecast distribution from the core VARX via entropic tilting. By extracting the marginal content of X_2 (projecting out X_1), one could drive the forecast errors down to $\hat{\varepsilon}$ rather than $\hat{u}$ from the core VARX.

To implement this idea, full re-estimation of the expanded VARX seems required in order to compute $\bar{B}_1$ and $\bar{B}_2$; however, the upcoming procedure and result will demonstrate that this not the case. The definition below describes a procedure to instead pin down the *difference* between Equations 3.4 and 3.5 as a post-estimation step or, equivalently, the excess terms in $\hat{u}$ beyond $\hat{\varepsilon}$. The steps of the procedure are driven by familiar results from least squares estimation, even though the model is estimated using Bayesian methods. This is because of the deep connection between OLS with dummy variables and the posterior mean in Bayesian estimation—a connection which traces back to Theil (1963) and has been exploited since the inception of the BVAR literature since Litterman (1980).

Definition 26. (Baseline Adjustment Procedure under Prior with Kronecker Structure) Consider the models and priors outlined in Assumption 12. Define the following procedure for adjusting the forecast distribution

⁸Both P and R are defined below in Definition 26,

of the core VARX based on external information X_2 .

1. (*Construct Inputs for Adjustment Procedure*): Given the prior for the full-information VARX as stated in [Definition 23](#) and reiterated in [Assumption 12](#), do a Cholesky decomposition of the portion Φ_0 , which characterizes the prior variance of the coefficients

$$\Phi_0 = LL' = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix}'$$

and retain $L_{11}^{-1}, L_{21}, L_{22}^{-1}$. Then build, in order,

$$\begin{aligned} X_d &= -L_{22}^{-1}L_{21}L_{11}^{-1} \\ Y_d &= X_d B_{0,1} + L_{22}^{-1}B_{0,2} \\ X_* &= \begin{pmatrix} X_{*,1} & X_{*,2} \end{pmatrix} = \begin{pmatrix} X_1 & X_2 \\ L_{11}^{-1} & 0 \\ X_d & L_{22}^{-1} \end{pmatrix} \\ V &= (X_1'X_1 + L_{11}^{-1'}L_{11}^{-1})^{-1} = (X_1'X_1 + \Phi_{0,11}^{-1})^{-1} \\ R &= -VX_d'(I_{K_2} + X_dVX_d')^{-1}[Y_d - X_d\bar{A}] \\ Y_* &= \begin{pmatrix} Y \\ L_{11}^{-1}B_{0,1} \\ X_dB_{0,1} + L_{22}^{-1}B_{0,2} \end{pmatrix} \\ \hat{u}_* &= Y_* - X_{*,1}(\bar{A} - R) \end{aligned}$$

This step effectively creates dummy observations in X_d and Y_d as well as an augmented sample in X_* and Y_* that encode the prior information.

2. (*Extract Marginal Variation from X_2*): Regress $X_{*,2}$ on $X_{*,1}$ to get coefficient matrix P and corresponding residuals, denoted $\hat{v}_*$ (including dummy observations), which capture marginal variation in X_2 not in X_1 , adjusted for shrinkage on X_1 's coefficients plus any prior correlation in coefficients across blocks.

3. (Model the Short Regression Errors Given External Info, under Shrinkage): Define

$$h(v) = vH$$

where H represents OLS coefficients from regressing $\hat{u}_*$ from Step 1 on $\hat{v}_*$ from Step 2:

$$H = (\hat{v}_*'\hat{v}_*)^{-1}\hat{v}_*'\hat{u}_*$$

4. (*Tilt*): Adjust draws from the baseline forecast distribution of the core VARX $f_A(y_{T+1:T+h}|Y)$ using entropic tilting under the constraint function and constraint:

$$g(y_{T+1:T+h}) = y_{T+1} \tag{3.6}$$

$$\bar{g} = x_{T+1,1}(\bar{A} - R) + h(x_{T+1,2} - x_{T+1,1}P) \tag{3.7}$$

where $(x_{T+1,1}, x_{T+1,2})$ are the first and second block of covariates for the next period, $\bar{A}$ the posterior mean, R from Step 1, P from Step 2, and H from Step 3.

Remark. (Leading Case of No Prior Correlation Between Blocks) Minnesota-like priors would typically assume no prior correlation between coefficients hitting the separate blocks X_1 and X_2 .⁹ Therefore, Φ_0 is block-diagonal

$$\Phi_0 = \begin{pmatrix} \Phi_{0,11} & 0 \\ 0 & \Phi_{0,22} \end{pmatrix}$$

This implies that L_{21} , X_d , and R all equal zero, which simplifies expressions in Step 1. I view this as the leading case. However, I do not assume this to allow for full generality.

The next result serves as justification for the procedure above. Rather than ad hoc, the adjusted forecast distribution from the above procedure exactly coincides with the forecast distribution one obtains under re-estimation, which I view as the full-information ideal.

Proposition 17. (re-estimation and Tilting Equivalence) *Suppose one follows the procedure of [Definition 26](#).*

⁹See ([Karlsson, 2013](#)) for details.

The mean of the tilted forecast distribution $f_*(y_{T+1:T+h})$ in the first period

$$\mathbb{E}_*[y_{T+1}] = x_{T+1,1}(\bar{A} - R) + (x_{T+1,2} - x_{T+1,1}P)H$$

equals the mean of the posterior predictive distribution of the full VAR from Proposition 15

$$\mathbb{E}_B[y_{T+1}|Y] = x_{T+1,1}\bar{B}_1 + x_{T+1,2}\bar{B}_2$$

Proof. See proof in C.3.3 for details. But to give an overview, the proof and intuition consists of exploiting four well-known results from estimation of linear models.

1. The forecast distributions of the core and full-information VARXs are built from the Bayesian posterior mean of the coefficients, but one can exploit the equivalence between Bayesian estimation of linear models and OLS with dummy variables. These results trace back to Theil (1963) and are well-known in the VAR literature since Litterman (1980). Then by recasting Bayesian posterior means as OLS estimators with an augmented sample, one can apply well-known results and formulas from least-squares estimation to characterize and bridge the gap between coefficients in Expressions 3.4 and 3.5.
2. The baseline forecast from the core VAR is constructed using $\bar{A}$, which may already incorporate the effect of information in X_2 that is correlated with X_1 . In other words, X_1 may already proxy for X_2 in an OVB sense, so that $\bar{A}$ encodes more than just the effect of X_1 . Standard OVB expressions, now with dummy variables, can pin that down.
3. The last point suggests that one must isolate the marginal information in X_2 , not already proxied for by X_1 . This calls for using the Frisch-Waugh-Lovell theorem to isolate the effect of marginal information on Y or, equivalently, the residuals u .
4. Finally, additional regressors require additional coefficients and terms in the prior. Priors for these additional coefficients can be implemented simply by more dummy observations. The effect of additional observations on OLS estimates is straightforward to compute. The expressions are similar in spirit to “leave-out” formulas as one would see in cross-validation, but they are “add-in” rather than “leave-out.”

By chaining together from these results, one can obtain a clear path working from the posterior mean of the core VARX to the posterior mean of the full-information VARX. This path for the posterior carries over to

the corresponding forecast distribution as well. $\square$

Before concluding this section, I now offer an intuitive description for each step of the procedure. These comments underpin the recommendation for the variations and extensions of [Section 3.3.2](#) and [Section 3.3.3](#).

1. *Step 1 as Sample Augmentation:* Step 1 augments the true sample from (Y, X_1, X_2) to $(Y_*, X_{*,1}, X_{*,2})$ with enough dummy observations to fully encode the prior of the implicit extended VARX. The objects computed in this step are precisely dummy variables needed to implement the prior in [Definition 23](#).

Similarly, Step 1 can also be thought of as augmenting the sample that was used to estimate $\bar{A}$ with another block of dummy variables to incorporate the additional coefficients for block X_2 . In fact, the posterior mean $\bar{A}$ in the core VARX under the prior of [Definition 22](#) can be calculated by regressing Y_* on $X_{*,1}$, but omitting the last block of regressors. Therefore, Step 1 computes additional objects to handle the expanded VARX and if the core VARX was estimated with dummy observations, one can simply append on additional observations to the estimation matrices.

2. *Step 2 Extracts the Appropriate “Marginal” Variation in X_2 :* Because X_1 may already proxy for X_2 in an OVB sense, subsequent adjustment of the forecast distribution should be based only on the *marginal* variation in X_2 . By the logic of the FWL theorem, this is done by residualizing X_2 in an appropriate sense, carried out in Step 2, with two adjustments made to the residualization because the model is estimated with a prior.

First, Step 2 step differs from residualization of plain OLS (without dummy observations) because X_1 was penalized in the core VARX to avoid overfitting, and so penalization should occur in this residualization step as well. Since the core VARX did not exploit all variation in X_1 , Step 2 should leave extra variation in $\hat{v}_*$. Unpenalized OLS would overfit and strip out too much variation from X_2 when producing $\hat{v}_*$. Practically, this is implemented by including the dummy variables encoding the prior during residualization. Analytically, in the leading case of Φ_0 being block diagonal, notice

$$P = (X'_{*,1}X_{*,1})^{-1}X'_{*,1}X_{*,2} = [X'_1X_1 + \Phi_{0,11}^{-1}]X'_1X_2$$

Recognize this as a shrinkage estimator that penalizes coefficients on X_1 , just as they were penalized in the core VARX by $\Phi_{0,11}$.

Second, an adjustment should be made for any prior correlation in coefficients between blocks, i.e. when Φ_0 is not block diagonal. The first block of coefficients in $\bar{A}$ have already been estimated; therefore, it

is necessary to make an adjustment for coefficients on the second block to account for correlation in the prior. Again, this ensures that one extracts the correct residual variation.

To see this, suppose X_1 and X_2 each consist of a single regressor so that Φ_0 is only two-dimensional. Because Φ_0 is positive definite, one can write

$$\Phi_0 = \begin{pmatrix} \phi_1^2 & \rho\phi_1\phi_2 \\ \rho\phi_1\phi_2 & \phi_2^2 \end{pmatrix} = LL' \quad \text{where} \quad L = \begin{pmatrix} L_{11} & L_{21}' \\ L_{21} & L_{22} \end{pmatrix} = \begin{pmatrix} \phi_1 & 0 \\ \rho\phi_2 & \sqrt{1-\rho^2}\phi_2 \end{pmatrix}$$

Using the definitions of $X_{*,1}$, $X_{*,2}$, and L , write out P in this special case

$$\begin{aligned} P &= (X_{*,1}'X_{*,1})^{-1}X_{*,1}'X_{*,2} \\ &= [X_1'X_1 + L_{11}^{-1'}L_{11}^{-1} + X_d'X_d]^{-1}[X_1'X_2 + X_d'L_{22}^{-1}] \\ &= \left[X_1'X_1 + \frac{1}{\phi_1^2(1-\rho^2)}\right]^{-1} \left[X_1'X_2 - \frac{1}{\phi_1^2(1-\rho^2)} \cdot \rho \frac{\phi_1}{\phi_2}\right] \end{aligned}$$

Notice that this now shrinks towards $-\rho\phi_1/\phi_2$. This centering changes the variation that enters into $\hat{v}_*$, which has an effect on the estimates in Step 3. Namely, if $\rho > 0$ or ϕ_2 is small, this will lead to a residual $\hat{v}_*$ that has “extra noise” than if $\rho = 0$. This extra noise will then attenuate the estimate in Step 3, and therefore help with properly accounting for the prior correlation between coefficient values.

As a summary of Step 2, the residual variation in $\hat{v}_*$ that is left for use in forecast adjustment depends upon both (1) how heavily coefficients on X_1 were penalized in the core VARX and (2) whether the coefficients of the first block are already accounting for much of the variation common to X_1 and X_2 .

3. *Step 3 Estimates the Marginal Impact:* Having extracted the marginal information, this step estimates the predictive power that marginal information has for the errors and outcomes of the core BVAR.

Consider the leading case where $\Phi_0 = \text{diag}(\Phi_{0,11}, \Phi_{0,22})$. The coefficients H which translate the marginal information in $\hat{v}_*$ into $\hat{u}_*$ satisfy

$$\begin{aligned} H &= (\hat{v}_*' \hat{v}_*)^{-1} \hat{v}_*' \hat{u}_* \\ \hat{v}_*' \hat{v}_* &= \underbrace{(X_2 - X_1 P)'(X_2 - X_1 P)}_{(i)} + \underbrace{P' \Phi_{0,11}^{-1} P + \Phi_{0,22}^{-1}}_{(ii)} \\ \hat{v}_*' \hat{u}_* &= \underbrace{(X_2 - X_1 P)'(Y - X_1 \bar{A})}_{(i)} + \underbrace{P' \Phi_{0,11}^{-1} (B_{0,1} - \bar{A}) + \Phi_{0,22}^{-1} B_{0,2}}_{(ii)} \end{aligned}$$

Notice that (i) represents what you would use without penalization, and these portions therefore capture how marginal information can be translated into updates for the short regression. Portion (ii) corresponds to the dummy variables and implements shrinkage. From (ii), notice the coefficients in H are shrunk towards

$$P'\Phi_{0,11}^{-1}(B_{0,1} - \bar{A}) + \Phi_{0,22}^{-1}B_{0,2}.$$

Therefore, the coefficients governing tilting are directed towards the prior $B_{0,2}$ plus an adjustment $B_{0,1} - \bar{A}$ to remove for any effect already captured by the first block.

4. *Step 4 Recenters while Preserving Uncertainty*: Lastly, an often desirable feature of tilting is that the point forecast changes as in conditioning, but the variance for outcomes used in the constraint does not collapse to zero, as it would for hard conditioning. Under hard conditioning $y_{T+1} = \bar{g}$, there would be no uncertainty for period $T + 1$ and reduced uncertainty for subsequent periods. This is a general feature of entropic tilting, extensively discussed in [Antolín-Díaz et al. \(2021\)](#) and exploited in other papers that utilize tilting for forecast updates like [Krüger et al. \(2017\)](#).

3.3.2 Variations on the Tilting Procedure

Relative to the re-estimation-equivalent tilting procedure of [Section 3.3.1](#), this section covers alternative procedures that depart from exact equivalence but may be preferable on the grounds of practicality or empirical performance.

And despite breaking exact re-estimation equivalence, the procedures of this section all build upon the main ideas of [Section 3.3.1](#) to define data-driven tilting procedures that exploit additional predictors similar to the re-estimation-equivalent steps. In particular, they all extract marginal information from additional covariates and model the core VARX errors to drive an adjustment. The steps remain the same, with only the details within steps changing.

As a first variation, the following procedure removes any restrictions on the model used to translate the extracted marginal information in X_2 into forecast adjustments. Specifically, Step 3 in [Definition 26](#) used a linear model, whereas the procedure below allows for any model. This flexibility may prove better at exploiting the marginal information in $\hat{v}_t$.

Definition 27. (General Procedure) Again consider the priors in [Assumption 12](#) for the estimated core VARX and implicit expanded BVARX. Further assume that $\Phi_{0,21} = 0$, i.e. independence across blocks,

which implies $R = 0$. Repeat the steps of [Definition 26](#), but with the following modification:

3. (*Model the Short Regression Errors Given External Info*): Given the residuals $\hat{v}_*$ and $\hat{u}_*$ including dummy variables, train a predictive model

$$\hat{u}_t = h(\hat{v}_t)$$

Unlike the previous procedure, this need not be linear regression and can be any predictive model.

Remark. (Use of Dummy Variables) Though the goal is no longer exact re-estimation equivalence with the implicit extended VARX, it is still useful to append dummy variables to the sample and use them in this modified procedure.

In Step 2, use of dummy variables was critical to extracting the correct marginal variation in $\hat{v}_*$ under appropriate penalization of X_1 , as discussed in the comments following [Definition 26](#). That requirement does not change, even when changing the model translating this marginal information into forecast adjustments.

In Step 3, the use of dummy variables also separately helped push the model towards the prior beliefs of the extended VARX, which remains useful even if one is not interested in exact equivalence. This helps with the two important tasks of (1) imposing shrinkage on parameters for the additional covariates and (2) centering on reasonable dynamics—both of which are useful regardless of the predictive model $h(\cdot)$. Further, the BVAR literature commonly adopts Minnesota or Minnesota-like priors; therefore, there is a reasonable consensus and strong intuition around this common, intuitive, well-understood prior. The inclusion of dummy variables effectively imports this to the generalized setting, which simplifies the task of prior specification for the otherwise unrestricted model of Step 3. In this way, the dummy variables serve as pseudo-observations in the sense of [Del Negro and Schorfheide \(2004\)](#), adding observations generated by dynamics “representative” of prior beliefs.¹⁰

Lastly, in crafting the dummy variables for this procedure, I assumed that $\Phi_{0,21} = 0$, while [Definition 26](#) allowed nonzero for full generality. This zero restriction generates dummy variables that encode independence across blocks in the implicit extended VARX. This is appropriate because the dummy variables are included to impose shrinkage and centering on reasonable dynamics, rather than enforce beliefs about a relationship between model parameters. Indeed, the alteration for Step 3 breaks formal equivalence to re-estimation of an expanded model, so that the coefficient for blocks X_1 and X_2 effectively forecast outcomes using two separate models. It is difficult to imagine a setting where the econometrician has well-informed beliefs about

¹⁰That said, if the econometrician is able to formulate and impose a reasonable prior in this modeling step that implements shrinkage towards reasonable dynamics, they could of course discard the dummy observations.

the relationship between parameters in the distinct models, and that the dummy variables would encode that correctly. In fact, even in traditional settings where one estimates a standard VARX, independence in the coefficients across regressors is a commonly made assumption. Assuming $\Phi_{0,21} = 0$ seems appropriate.

As an alternative relaxation of [Definition 26](#), I next define a procedure that dispenses with the assumption of Kronecker structure in the prior variance for the core VARX. This is useful for the reasons discussed preceding [Definition 24](#).

Assumption 13. (Non-Kronecker Priors for Core and Expanded VARX) Suppose that one has estimated the core VARX using the prior of [Definition 24](#), given econometrician-specified values for the tuning parameters $(A_0, \Omega_{0,A}, S_0, \alpha_0)$.

Further suppose that one has access to additional covariates X_2 . These can be used to define an extended VARX that employs the prior of [Definition 24](#). Assume $B_{0,1} = A_0$, $\Omega_{0,11} = \Omega_{0,A}$, as well as identical S_0 and α_0 . The values for the remaining tuning parameters $(B_{0,2}, \Omega_{0,22}, \Omega_{21})$ can be chosen freely by the econometrician, as they would choose them if re-estimating the expanded VARX, subject to ensuring Ω_0 is symmetric positive definite.

The remark following [Assumption 12](#) on identical S_0, α_0 between priors also applies here. The assumption of $\Omega_{21} = 0$ is also done for simplicity, and this is common in practice.

Proposition 18. (Corresponding Posterior and Posterior Predictive Distribution) *Given the prior of [Assumption 13](#), the posterior distribution $p_A(A, \Sigma_A|Y)$ is nonstandard; however, draws from the posterior, $\{A^{(d)}, \Sigma_A^{(d)}\}_{d=1}^D$, can be generated by a straightforward Gibbs sampling scheme. Because the posterior of the implicit extended VARX has the same structure, an identical statement applies to $p_B(B, \Sigma_B|Y)$.*

The posterior predictive (forecast) distribution of the core VARX

$$f_A(y_{T+1}|Y) = \int f_A(y_{T+1}|A, \Sigma_A, Y) p_A(A, \Sigma_A|Y) dA d\Sigma_A$$

can be approximated using the posterior draws.

Proof. For the Gibbs sampling algorithm, see [Karlsson \(2013\)](#), Algorithm 2 (page 811). □

Because there are not closed form expressions for the posterior distributions $p_A(A, \Sigma_A|Y)$ and $p_B(B, \Sigma_B|Y)$, I cannot compare the posterior and posterior predictive (forecast) distributions analytically under this prior structure. However, one can define an adjustment procedure similar to [Definition 26](#) which has a clear, intuitive interpretation in terms of the non-analytical posteriors.

In the adjustment procedure below, I assume that one has available draws $\{a^{(d)}, \Sigma_A^{(d)}\}_{d=1}^D$, of the coefficients (a, Σ_A) from the posterior of the core VARX, as well as the conditional posterior mean of coefficients $\{\bar{a}^{(d)} | \Sigma_A^{(d)}\}_{d=1}^D$, corresponding to each $\Sigma^{(d)}$ draw, which is necessary to compute when sampling the posterior and cheap to store. Recall that $a = \text{vec}(A)$.

Definition 28. Take as given the models and priors outlined in [Assumption 13](#). As a pre-processing step, do a Cholesky decomposition of the assumed block-diagonal prior variance of the extended VARX

$$\Omega_0 = LL' = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix}',$$

retain $L_{11}^{-1}, L_{21}, L_{22}^{-1}$, and construct

$$\begin{aligned} X_d &= -L_{22}^{-1} L_{21} L_{11}^{-1} \\ y_d &= X_d \beta_{0,1} + L_{22}^{-1} \beta_{0,2}. \end{aligned}$$

Next, suppose that one has draws of the error variance $\{\Sigma_A^{(d)}\}_{d=1}^D$ from the posterior of the core VARX, as well as the associated conditional posterior $\bar{A}^{(d)} | \Sigma_A = \Sigma^{(d)}$ used to generate the draw of $A^{(d)}$ given $\Sigma^{(d)}$. Then for each pair $\Sigma_A^{(d)}, \bar{A}^{(d)}$, we essentially repeat the steps of [Definition 26](#), but modified to allow for the less restrictive prior.

1. (*Construct Inputs for Adjustment Procedure*): Build,

$$\begin{aligned} X_* &= \begin{pmatrix} X_{*,1} & X_{*,2} \end{pmatrix} = \begin{pmatrix} (\Sigma_A^{(d)})^{-1/2} \otimes X_1 & (\Sigma_A^{(d)})^{-1/2} \otimes X_2 \\ L_{11}^{-1} & 0 \\ X_d & L_{22}^{-1} \end{pmatrix} \\ V &= ((\Sigma_A^{-1} \otimes X_1' X_1) + L_{11}^{-1'} L_{11}^{-1})^{-1} = ((\Sigma_A^{-1} \otimes X_1' X_1) + \Omega_{0,A}^{-1})^{-1} \\ R &= -V X_d' (I_{K_2} + X_d V X_d')^{-1} [y_d - X_d \bar{a}^{(d)}] \\ y_* &= \begin{pmatrix} \text{vec}(Y(\Sigma_A^{(d)})^{-1/2}) \\ L_{11}^{-1} \beta_{0,1} \\ L_{22}^{-1} \beta_{0,2} \end{pmatrix} \\ \hat{u}_* &= y_* - X_{*,1}(\bar{a}^{(d)} - R) \end{aligned}$$

2. (Extract Marginal Variation from X_2): Identical instructions as [Definition 26](#).
3. (Model the Short Regression Errors Given External Info, under Shrinkage): Identical instructions as [Definition 26](#).
4. (*Compute for Tilting*): Compute

$$g^{(d)} = (I_N \otimes x_{T+1,1})(\bar{a}^{(d)} - R) + h((I_N \otimes x_{T+1,2}) - (I_N \otimes x_{T+1,1})P)$$

where $(x_{T+1,1}, x_{T+1,2})$ are the first and second block of covariates for the next period, $\bar{a}^{(d)}$ the (conditional) posterior mean of coefficients used in generating $a^{(d)}$, R is from Step 1, P from Step 2, and H from Step 3.

After repeating the above procedure looping over all draws, use entropic tilting under the following constraint function and constraint:

$$g(y_{T+1:T+h}) = y_{T+1} \tag{3.8}$$

$$\bar{g} = \frac{1}{D} \sum_{d=1}^D g^{(d)} \tag{3.9}$$

Take the resulting tilted forecast distribution as the adjusted forecast distribution.

Ignoring the loop over Σ_A draws, the procedure of [Definition 28](#) looks much like that of [Definition 26](#), which is no coincidence. Conditional on Σ_A , the prior on coefficients is still conjugate, even though the variance no longer has Kronecker form, and can be implemented with dummy variables. This dummy variable implementation underpinned the procedure of [Definition 26](#), and underpins this procedure as well. The only difference is that one must stack and transform the observations slightly differently, now that we have dispensed with the convenient Kronecker assumption the prior variance of coefficients.¹¹ And although this similar structure does not carry over to exact re-estimation equivalence, it does imply an intuitive result summarized in the following proposition.

Proposition 19. *Ignoring Monte Carlo approximation error, the procedure of [Definition 28](#) generates a forecast distribution $f_*(y_{T+1:T+h})$ satisfying*

$$\mathbb{E}_*[y_{T+1}] = \int w(\Sigma) \cdot \mathbb{E}_B[y_{T+1} | \Sigma_B = \Sigma, Y] \cdot p_B(\Sigma | Y) d\Sigma$$

¹¹See [Appendix C.2.1](#) for precise results and further discussion.

where

$$\begin{aligned}\mathbb{E}_B[y_{T+1}|\Sigma_B = \Sigma, Y] &= \int y_{T+1} f_B(y_{T+1}|B, \Sigma, Y) p_B(B|\Sigma, Y) dy_{T+1} \\ w(\Sigma) &= \frac{p_A(\Sigma|Y)}{p_B(\Sigma|Y)}\end{aligned}$$

Remark. (Interpretation) If $w(\Sigma) = 1$, the forecast distribution would be the same as that of the implicit extended VARX, marginalizing over posterior parameter uncertainty, just like Proposition 17. However, because the adjustment procedure reuses the draws Σ_A , it is effectively drawing from the the wrong distribution for the error variance. This induces a weighting function $w(\Sigma)$ that overweights coefficients consistent with the portion of the joint parameter space corresponding to larger values of Σ , versus the true marginal posterior $p_B(\Sigma|Y, X_1, X_2)$, and underweights smaller.¹²

Remark. (Use of $\bar{a}^{(d)}$ versus $a^{(d)}$) In Definition 28, I use $\bar{a}^{(d)}$ when doing the forecast adjustment procedure, which is the posterior mean of the coefficients under Σ_A , rather than the draw $a^{(d)}$, which satisfies

$$a^{(d)} = \bar{a}^{(d)} + (\bar{\Omega}_A^{(d)})^{-1/2} w \quad w \sim \mathcal{N}(0, I_{NK_1})$$

for some independent random normal variable w . If we applied the procedure with $a^{(d)}$, we would introduce a source of noise into the adjustment procedure, which would not average out to zero and would not offer the relatively clean and intuitive result given above.

Proof. Proposition 17 first established a path between posterior means

$$\bar{A} = \mathbb{E}_A[A|Y] \quad \rightarrow \quad \bar{B} = \mathbb{E}_B[B|Y]$$

via the procedure of Definition 26. I then mapped this into implications for the forecast distributions. Note that because of the convenient Kronecker structure in both the prior and posterior, there is a direct path between *marginal* posterior means, after marginalizing out uncertainty about Σ_A, Σ_B from each posterior distribution. See Corollary 5 in Appendix C.2.2 for more.

In contrast, after dispensing with the Kronecker structure, the steps of the new procedure establishes a

¹²This assumes that the additional covariates do, in fact, reduce the error variance.

map between *conditional* posterior means

$$\mathbb{E}_A[a|\Sigma_A = \Sigma, Y] = \mathbb{E}_A[\text{vec}(A)|\Sigma_A = \Sigma, Y] \quad \rightarrow \quad \mathbb{E}_B[b|\Sigma_B = \Sigma, Y] = \mathbb{E}_B[\text{vec}(B)|\Sigma_B = \Sigma, Y]$$

and the corresponding posterior predictive (forecast) distribution. More precisely $g^{(d)}$ in Step 4 of the procedure in [Definition 28](#) satisfies

$$g^{(d)} = x_{T+1} \mathbb{E}_B[b|\Sigma_B = \Sigma, Y] = \mathbb{E}_B[y_{T+1}|\Sigma_B = \Sigma, Y]$$

Then, since this procedure uses the draws of $\Sigma_A^{(d)}$, note

$$\bar{g} = \frac{1}{D} \sum_{d=1}^D g^{(d)} \approx \int \mathbb{E}_B[y_{T+1}|\Sigma_B = \Sigma, Y] p_A(\Sigma|Y) d\Sigma$$

One can rewrite to introduce $w(\Sigma)$ in the integrand, and this will equal the form stated in the proposition. Therefore, the procedure recenters the forecast distribution to match this $\bar{g}$. $\square$

3.3.3 Longer Horizon Forecasts

While I have focused on the adjustment to one-step-ahead forecasts, any of the procedures will (1) indirectly influence longer-horizon forecasts and (2) can be modified if one prefers for more direct modification of longer-horizon forecasts.

Starting with the second point, the procedure can be immediately applied to direct forecasts of longer horizons. In direct forecasting, the econometrician is interested in forecasting h -periods ahead and regresses y_{t+h} directly on the information available at time t .¹³ In this context, rather than the model of [Definition 18](#), one would directly estimate the h -step forecasting model

$$y_{t+h} = A_1 y_t + \cdots + A_p y_{t-p+1} + C z_t + u_t$$

This also stacks into the same form

$$Y = X_1 A + u$$

but with Y now containing h period ahead observations given the covariates and lags in X_1 . All results and

¹³See [Marcellino, Stock and Watson \(2006\)](#) for a comparison of direct versus iterated forecasts.

conclusions are then identical. If the econometrician is interested in direct forecasts of multiple periods, the procedure could be applied for each direct forecast.

Often however, forecasters instead use iterated forecast distributions from a single estimated VAR. I briefly state how this paper covers that setting.

First, tilting to hit a constraint one period ahead will indirectly adjust longer horizon forecasts as well. As described in the implementation remark following [Definition 25](#), tilting reweights draws of forecast *paths* $\{y_{T+1:T+h}^{(d)}\}_{d=1}^D$ to satisfy the given constraint. Though the constraint is only imposed on the initial period, the reweighting applies to the full path. As a result, tilting mimics conditional forecasting, which is further detailed in [Antolín-Díaz et al. \(2021\)](#), and the effect of the adjustment will propagate forward.¹⁴

Second, if one is concerned that this is still insufficient, one can add additional constraints for the tilting problem. Following the logic of direct forecasting and this paper, one can compute multi-step ahead residuals $\hat{u}_{t+s|t} = y_{t+s} - \hat{y}_{t+s|t}$ where $\hat{y}_{t+s|t}$ is the iterated forecast of y_{t+s} from the core VAR constructed at time t , for some horizon s . Repeat the procedure of [Definition 26](#) or some variant to determine the adjustment that should be made to multi-step ahead forecast errors based on the residual information in X_2 . Given this, one can then construct and append an additional constraint for s periods ahead to the tilting problem.

3.4 Empirical Application

This section illustrates and evaluates the procedures of [Section 3.3](#) via a nowcasting application, borrowed from [Krüger et al. \(2017\)](#). In this section, the core model is a quarterly BVAR consisting of a small number of macroeconomic aggregates commonly used for structural or policy analysis. Higher frequency monthly signals constitute the additional regressors driving forecast adjustment. Given these, [Krüger et al. \(2017\)](#) use the higher frequency data to estimate four auxiliary models and fully tilt the forecast distribution of the quarterly BVAR towards their one-step-ahead forecasts. In contrast, I will use the same data to extract the marginal information from the additional data and follow the procedures in [Section 3.3](#) to update the forecast distribution. I show that my recommended procedures perform comparably to fully tilting towards nowcasts as in [Krüger et al. \(2017\)](#).

¹⁴More precisely, if the starting distribution is normal and the tilting constraint a linear function of $y_{T+1:T+h}$ (covering forecasts from VARs conditional on parameters), the mean of the tilted distributions equals the mean of the conditional forecast distribution. When accounting for parameter uncertainty, this precise result breaks but the broad similarity between tilting and conditional forecasting remains.

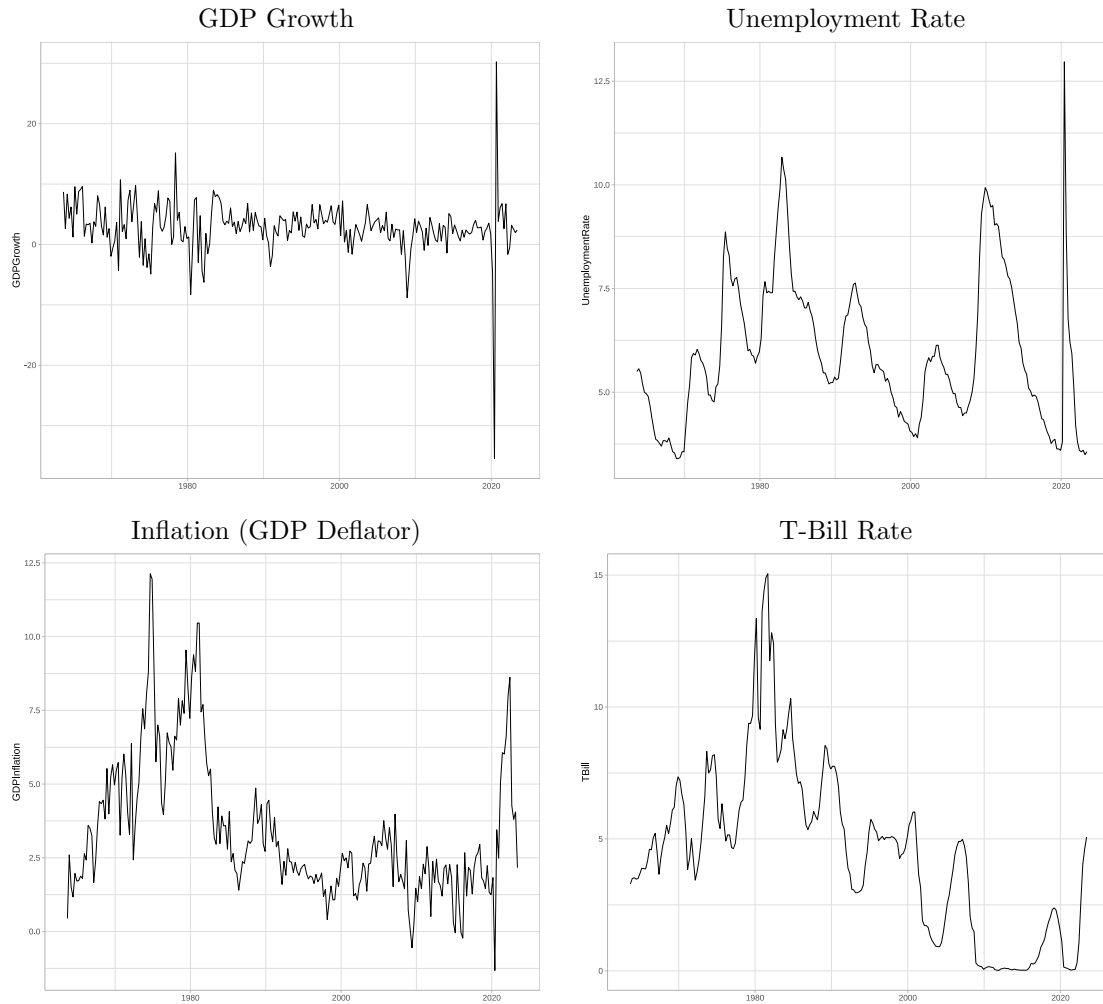


Figure 3.1: Time series used in the core BVAR

3.4.1 Data

This section describes the four core macroeconomic time series of interest that enter into the core quarterly BVAR, as well as the auxiliary data available used for forecast adjustment. Because this is a nowcasting application involving mixed frequencies and time series subject to revisions, I also describe the timing assumptions and use of real-time data for the exercise. I collect and prepare the series essentially as in [Krüger et al. \(2017\)](#) with one exception.¹⁵ Beyond that, I employ slightly different sample periods and modeling assumptions than [Krüger et al. \(2017\)](#); however, this section is self-contained and provides a full description of the sample and modeling assumptions I employ.

¹⁵Additional details on the data are given in [Appendix C.5.1](#).

Core Time Series The core time series of interest consist of quarterly observations for four outcomes: annualized quarterly GDP growth, annualized quarterly inflation as measured by the GDP Deflator, the unemployment rate, and the annualized interest rate on the three-month T-Bill. Figure 3.1 plots the series. See Appendix C.5 for more details on the exact construction of these four time series, which match the outcomes studied in Krüger et al. (2017).

These outcomes are representative of the set of time series often included in core macroeconomic forecast models, and well-suited to the setting of this paper. For example, a small-scale VAR with these outcomes can be used for dynamic structural analysis, as has been done since Sims (1980). At the same time, these outcomes are important ingredients and metrics for policymaking. That is because these core outcomes span different sectors or “blocks” of the economy. Therefore, having a stable model for the joint dynamics of these core time series will likely always be of interest for analysis. At the same time, the procedures of Krüger et al. (2017) and this paper can be layered on top to improve forecast accuracy using arbitrary predictive signals, without cluttering the simple core model with further series that provide purely predictive insight rather than economic insight.

Auxiliary Data and Timing Assumptions The out-of-sample exercise follows Krüger et al. (2017) and imagines nowcasting outcomes for quarter t in the core model, using additional monthly information through the first month of quarter t . Rather than construct a forecast using only data up to quarter $t - 1$, the econometrician can update the forecast based on additional higher frequency information.

Therefore, for each outcome in the core model, Krüger et al. (2017) identifies potentially informative time series for each outcome of interest in the core model, and I will also use these to test my procedure. By design of the out-of-sample exercise, these additional series are collected through the first month of quarter t and could be used in real time for forecast improvement.¹⁶ Table 3.1 lists the additional monthly predictors, while Appendix C.5.1 gives further details on their definition and construction. The table groups together the monthly predictors by the core outcome they are meant to help forecast. Krüger et al. (2017) use the grouped data to construct and estimate four separate nowcast models, using only the monthly data series “matched” with the corresponding outcome. In contrast, my procedure will extract and exploit the marginal information in all additional predictors in one step to update the forecasts of the four outcomes simultaneously.

¹⁶To be precise, as stated, the latest observation used is for the first month in quarter t . This is typically released in the second month of quarter t , which is the time of forecast update in this exercise.

Outcomes	Additional Predictors	Available Sample
GDP Growth	Employment Growth ISM Production Index, Manufacturing	(1947-06, 2023-06) (1947-02, 2023-07) (1948-01, 2022-03)
GDP Inflation	CPI ex-Food and Energy, CPI for Food CPI for Energy PPI: Private Capital Equipment Deflator, New Single-Family Houses	(1947-06, 2023-06) (1957-02, 2023-07) (1947-02, 2023-07) (1957-02, 2023-07) (1947-05, 2023-07) (1964-02, 2023-07)
Unemployment Rate	Employment Growth New claims for unemployment insurance	(1948-03, 2023-06) (1947-02, 2023-07) (1967-02, 2023-07)
3-Month T-Bill Rate	3-Month T-Bill Rate at Mid-Month 6-Month T-Bill Rate at Mid-Month	(1934-03, 2023-06) (1954-02, 2023-07) (1959-01, 2023-07)

Table 3.1: Additional predictors and associated outcomes they are intended to help predict. All outcomes are quarterly frequency, and the predictors are all monthly frequency. The available sample is partially dependent upon the data sample and extraction data; see [Appendix C.5.1](#).

Sample and Evaluation Period Owing to the data availability stated in [Table 3.1](#), I use data back 1970:Q1 for estimating the models, and I start evaluating forecast accuracy in 1992:Q1 until 2019:Q4 (inclusive) to abstract from the sharp outliers due to COVID. As I work my way through the sample period, I also estimate the model following an expanding-window convention common to macroeconomic modeling, rather than estimate on a fixed window. In [Appendix C.5.3](#), I also examine accuracy using a shorter sample and over subsamples. The results and conclusions do not qualitatively change over subsamples or later time periods.

Even though the time series in the core BVAR are available in prior quarters prior to 1970:Q1, I limit to the shorter sample for two reasons. First, this avoids the issue of differing sample sizes between core model and the two post-estimation adjustment methods that exploit auxiliary information. Therefore, the only difference between forecasts from the BVAR, my procedure, and nowcasting are the number of series used and the modeling approach, not sample. Moreover, under this restriction, my procedure will also exactly coincide with (or mimic, depending on the prior) estimation of a VARX that could feasibly be estimated on this restricted sample period. Second, since [Primiceri \(2005\)](#), time-varying models have been increasingly employed to accommodate the apparent change in dynamics during the postwar period for the United States macroeconomy. However, I wish to use a fixed coefficient model for simplicity, connection to the analytical results, and clarity of exposition. This recommends using a shorter sample over which dynamics have been

more stable.

These are choices of convenience and should not be interpreted as restrictions on the adjustment method of this paper. I aim only to give a sharp comparison during a relatively homogeneous sample. In practice, given a longer sample and outliers due to COVID or other extreme events, one could, for example, estimate a time-varying model as in [Carriero, Clark, Marcellino and Mertens \(2022\)](#) and follow the general adjustment procedure of [Definition 27](#). The intuition of regressing errors on the marginal information, under penalization, still stands and provides useful guidance about how to tilt the forecast distribution in a data- and re-estimation-driven manner.

Real-Time Data Data revisions are an important feature of many macroeconomic time series; therefore, I use real time data where available, as in [Krüger et al. \(2017\)](#). This includes using real-time data for all the core time series of interest,¹⁷ as well as real-time data for the monthly unemployment rate and employment series. For other auxiliary time-series where real-time data are not available, I uses the time series history available as of December 2023.

3.4.2 Modeling Specifications

In this section, I outline the forecasting models and corresponding prior assumptions. All the models are linear models for time series data, estimated using Bayesian methods. Therefore, I rely heavily on priors following the Minnesota prior of [Litterman \(1980\)](#), but also allowing for estimation of the forecast error variance. It is common in the literature and can also be used to define priors exactly matching the forms of [Definition 22](#) and [Definition 24](#), under which I proved the analytical results of [Section 3.3](#).

To first state these prior assumptions in general, I use the notation (B, Σ) to denote the generic coefficient matrix and forecast error variance, while $\beta = \text{vec}(B)$. Also, I also use Φ_ℓ to denote the ℓ th lag matrix in the VAR portion of the model. Refer to [Definition 18](#) or [Definition 20](#) for the exact stacking I employ. Again, (B, Σ) are generic in this section and could refer to either A or B in [Section 3.2](#).

Assumption 14. (Minnesota Assumptions for Prior Mean of Coefficients) In setting the prior mean of the coefficients in B (or, equivalently, $\beta = \text{vec}(B)$), assume that for any regressor $k \in \{1, \dots, K\}$ and $i \in \{1, \dots, N\}$

$$\mathbb{E}[B_{ki}] = \mathbb{E}[\beta_{0, K \cdot (i-1) + k}] = \mathbf{1}\{k = i\} \cdot \mathbf{1}\{y_{t,i} \not\sim I(0)\}.$$

¹⁷Note that the T-Bill rate is not revised and does not require real-time data.

β_0 is mostly a vector of zeros except (potentially) for entries that correspond to an outcome's first own lag, in which case the entry equals 1 if the series is not well-modeled by an $I(0)$.

Assumption 15. (Minnesota Assumptions for Prior Variance of Coefficients) Suppose that the full prior for model parameters requires specifying $p(\beta) \sim \mathcal{N}(\beta_0, \Omega_0)$, as in [Definition 24](#). Let σ_i^2 denote the variance of residuals from a univariate AR(p) model estimated on the i th variable in the VAR, using the full sample, where p matches the VAR order of the VAR. Then specify Ω_0 as a diagonal matrix, with the components of the diagonal set as follows, using the σ_i^2 as well as econometrician-specified tuning parameters $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$:

1. (*Own Lags*): The prior variance of coefficients corresponding to own lags for endogenous variables satisfy—for any equation $i = 1, \dots, N$ and lag $\ell = 1, \dots, p$ — the following:

$$\text{Var}(\beta_{K \cdot (i-1) + N \cdot (\ell-1) + i}) = \text{Var}(B_{N \cdot (\ell-1) + i, i}) = \text{Var}(\Phi_{\ell, ii}) = \left(\frac{\lambda_1}{\ell^{\lambda_3}} \right)^2$$

2. (*Cross-Lags/Regressors*): The prior variance for a coefficient in the i th equation hitting another outcome $j \neq i$ at lag ℓ are as follows:

$$\text{Var}(\beta_{K \cdot (i-1) + N \cdot (\ell-1) + j}) = \text{Var}(B_{N \cdot (\ell-1) + j, i}) = \text{Var}(\Phi_{\ell, ij}) = \lambda_2^2 \left(\frac{\sigma_i}{\sigma_j} \right)^2 \left(\frac{\lambda_1}{\ell^{\lambda_3}} \right)^2$$

Mechanically, this allows for different shrinkage across equations.

3. (*Deterministic Regressors*): The prior variance for a coefficient in the i th equation hitting the m th deterministic regressor (e.g. a constant, a trend, seasonality dummies, etc.) for any $m \in \{1, \dots, M\}$ is

$$\text{Var}(\beta_{K \cdot (i-1) + Np + m}) = \text{Var}(B_{Np + m, i}) = \text{Var}(C_{im}) = \sigma_i^2 (\lambda_1 \lambda_4)^2 \quad m = 1, \dots, M$$

The tuning parameter λ_4 is typically taken as large to express a diffuse prior.

Reasonable tuning parameter values from the literature are $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.2, 0.2, 1, 1000)$. Notice λ_1 controls overall shrinkage, λ_2 controls cross-variable shrinkage, λ_3 controls decay in the prior variance with increasing lag, and λ_4 controls shrinkage on deterministic regressors.

Assumption 16. (“Minnesota-Like” Assumptions for Prior Variance of Coefficients) Suppose that the full prior for model parameters requires specifying $p(\beta|\Sigma) \sim \mathcal{N}(\beta_0, \Sigma \otimes \Phi_0)$, as in [Definition 22](#), where Φ_0 is a $K \times K$ matrix in which the k th diagonal entry specifies/controls a common variance component applied to

the k th regressor in each equation. Again, let σ_i^2 denote the variance of residuals from a univariate AR(p) model estimated on the i th variable in the VAR, using the full sample, where p matches the VAR order.

To implement a “Minnesota-like” prior, specify Φ_0 as a diagonal matrix with the components of the diagonal set as follows, using the σ_i^2 as well as econometrician-specified tuning parameters $(\lambda_1, \lambda_3, \lambda_4)$.

1. (*Own-Lags and Cross-Lags/Regressors*): For any endogenous outcome/regressor $j \in \{1, \dots, N\}$ and lag $\ell \in \{1, \dots, p\}$, let

$$\text{diag}(\Phi_0)_{N \cdot (\ell-1) + j} = \left(\frac{1}{\sigma_j^2} \right) \left(\frac{\lambda_1}{\ell \lambda_3} \right)^2$$

2. (*Deterministic Regressors*): For any $m = 1, \dots, M$, let

$$\text{diag}(\Phi_0)_{Np+m} = (\lambda_1 \lambda_4)^2$$

This is similar to Assumption 15, except that this prior does not shrink coefficients on cross-lags more. Mechanically, there is no longer $\lambda_2 < 1$ to implement stronger shrinkage for cross-lags.

Assumption 17. (Diffuse Prior Assumptions for Σ) Suppose that the full prior for model parameters requires specifying $p(\Sigma) \sim IW(S_0, \alpha_0)$. The prior can be set in a diffuse manner by taking

$$\alpha_0 = N + 2 \quad S_0 = (\alpha_0 - N - 1) \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$$

In effect, this choice takes α_0 to be the smallest possible value such that the prior mean $S_0/(\alpha_0 - N - 1)$ for Σ still exists, and then sets S_0 so that the prior mean of Σ is centered on $\text{diag}(\sigma_1^2, \dots, \sigma_N^2)$, which is a hard assumption in the original Minnesota prior of Litterman (1980), but only a centering in this case.

Core BVAR Using the $N = 4$ core time series described in Section 3.4.1, I estimate a quarterly fixed-coefficient core BVAR with $p = 3$ lags and a constant, which is nested by the VARX of Definition 18. For the prior, I consider two prior specifications.

The first prior specification exactly matches the form $p(A|\Sigma_A)p(\Sigma_A)$ in Definition 22; therefore, it is conjugate with Kronecker structure in the (prior and posterior) variance of coefficients, which permits re-estimation equivalence. To fully specify $p(A|\Sigma_A)p(\Sigma_A) \sim \mathcal{N}(A_0, \Sigma_A \otimes \Phi_{0,A})IW(S_0, \alpha_0)$, I set

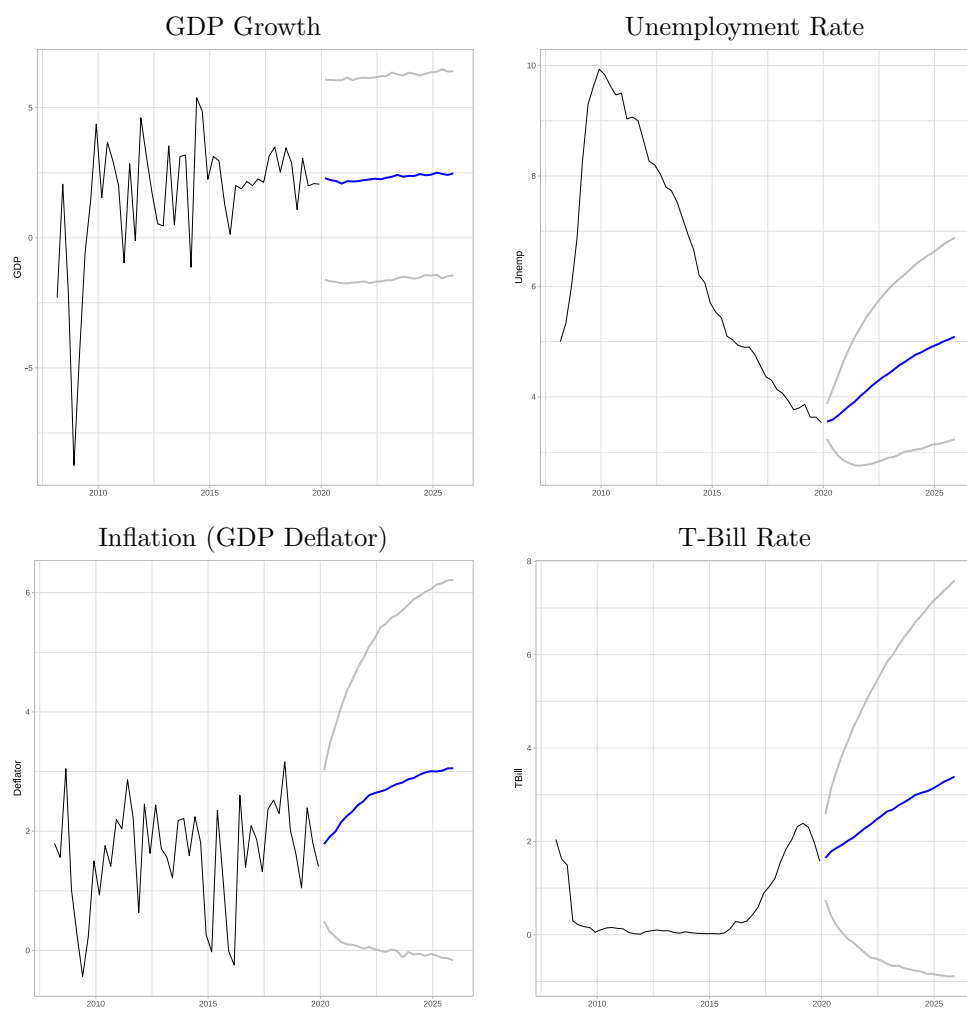


Figure 3.2: Example forecasts from the Core BVAR, under the second (non-Kronecker) prior specification. The blue line represents the median forecast, while the gray lines represent the 10th and 90th quantiles of the forecast distribution.

- A_0 as in Assumption 14, with the first own lag for GDP Growth and inflation centered on zero, and the first own lag for unemployment and the T-Bill rate centered on one.
- $\Phi_{0,A}$ as in Assumption 16.
- (S_0, α_0) as in Assumption 17.

The second prior specification instead matches the form of $p(A)p(\Sigma_A)$ in Definition 24; therefore, it is conjugate, but lacks the Kronecker structure in the variance of coefficients. To fully specify $p(A)p(\Sigma_A) \sim \mathcal{N}(A_0, \Omega_{0,A})IW(S_0, \alpha_0)$, I set

- A_0 as in Assumption 14, with the first own lag for GDP Growth and inflation centered on zero, and the first own lag for unemployment and the T-Bill rate centered on one.
- $\Omega_{0,A}$ as in Assumption 15.
- (S_0, α_0) as in Assumption 17.

As discussed in Section 3.3.2, this second specification will break exact re-estimation equivalence of my procedure, but it allows me to test the improvement one might obtain by departing from exact equivalence while still maintaining the main spirit of the adjustment procedure.

Remark. (Specification of Core Model and its Prior) Rather than craft the best possible quarterly forecasting model, my goal in this section is only to establish a reasonable baseline model, so that I can then compare alternative post-estimation adjustment procedures. The same statement applies to prior specification. In my comparison, the core model is held fixed, regardless of the post-estimation procedure. Therefore, I chose a model and prior specifications matching Definition 22 and Definition 24 to tie back directly to the analytical results of the previous section.

I highlight this because the preceding literature in Krüger et al. (2017) and Tallman and Zaman (2020) adopted richer time-varying specifications (e.g. Primiceri (2005)) as the core model; however, their comparison of interest was between baseline forecasts from the core model and adjusted forecasts. Their work and the broader nowcasting literature have already convincingly argued that incorporating additional higher-frequency information improves forecast accuracy relative to baseline forecasts from a lower frequency VAR (Bańbura, Giannone, Modugno and Reichlin, 2013). My interest in this paper and the empirical exercise is instead the manner of incorporation and identifying whether my proposed tilting procedure is deficient relative to full nowcasting as in Krüger et al. (2017).

A similar statement applies to choice of prior, where I again adopt a simple prior specification. The literature has developed popular additions to and extensions of the Minnesota prior, e.g. the “sum of coefficients” extension of [Doan, Litterman and Sims \(1984\)](#). The literature has also developed alternative methods for setting the tuning parameters of a given prior based on marginal likelihood maximization (as in [Del Negro and Schorfheide \(2004\)](#) and [Carriero, Clark and Marcellino \(2015a\)](#)) or hierarchical Bayesian approaches ([Giannone et al., 2015](#)). These are all worth using in the BVAR, but I omit them from analysis for the same reasons.

That said, a simple extension that still fits the form [Definition 24](#) would be to exploit the grouping implicit in [Table 3.1](#). Namely, adopt the second prior specification above, but for the GDP equation, apply further shrinkage to coefficients hitting predictors not selected to help forecast GDP. A similar statement would apply to further outcomes.

Nowcasts [Krüger et al. \(2017\)](#) use the additional data to define and estimate four auxiliary models—one for each outcome in the core BVAR—which I now describe. These are full-fledged models that require their own set of modeling and prior assumptions used to predict a given quarterly observation, which I will describe. See also [Appendix C.5.2](#) for figures that depict parts of the nowcasts’ construction.

To predict one-step-ahead GDP growth, [Krüger et al. \(2017\)](#) follow [Carriero, Clark and Marcellino \(2015b\)](#) and regress quarterly GDP growth for quarter t , denoted y_t , on a set of regressors X_t available for predicting GDP growth. This can be written as the linear model

$$y_t = X_t' \gamma + v_t \quad v_t \sim \mathcal{N}(0, \sigma^2). \quad (3.10)$$

The set of predictors collected in X_t are the lag of quarterly GDP growth (i.e. from quarter $t - 1$), the two additional monthly predictors for GDP growth (observed in the first month of quarter t) listed in [Table 3.1](#), plus a constant.¹⁸ Notice that this model is exactly a VARX for GDP growth with one lag, two additional nondeterministic regressors, and a constant. As a result, [Krüger et al. \(2017\)](#) implement a Minnesota-type prior on the coefficients in γ . Specifically, they employ a non-Kronecker prior like [Definition 24](#), $p(\gamma)p(\sigma^2) = \mathcal{N}(\gamma_0, \Omega_{0,\gamma})IW(S_0, \alpha_0)$. They set γ_0 following [Assumption 14](#), $\Omega_{0,\gamma}$ following [Assumption 15](#), and S_0 and α_0 as in [Assumption 17](#).

Note that [Carriero et al. \(2015b\)](#) also explores specifications where the model in [Expression 3.10](#) has

¹⁸Because GDP is subject to (sometimes substantial) revisions, I also use real-time data in the estimation of and forecasting steps for GDP growth, since they are readily available.

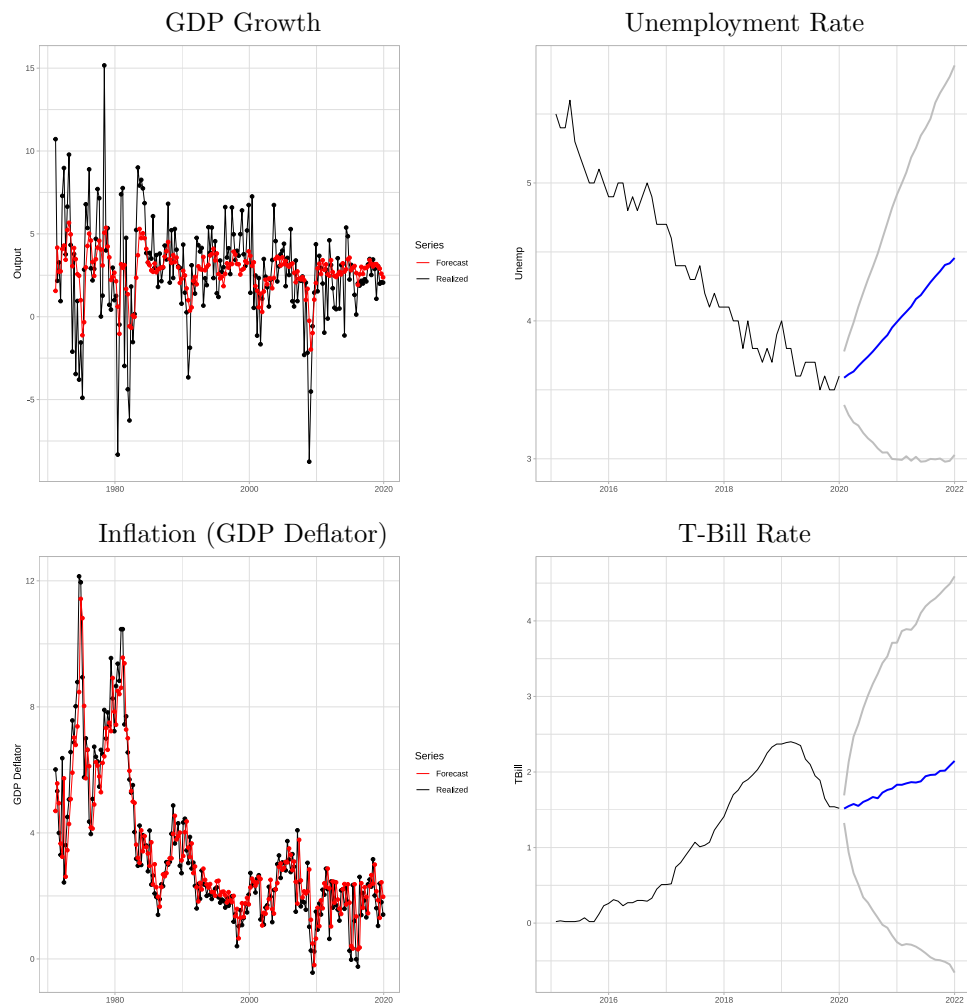


Figure 3.3: Example of nowcast construction. In the lefthand column, the black line represents the realized time series. The red line represents the fitted values from the regression used to generate the respective nowcast. The righthand column shows the history and forecast for the monthly unemployment rate and TBill. The final observation and first two forecasts are averaged to product the nowcast.

stochastic volatility, which is especially helpful for long macro samples. For simplicity, I omit this from testing, but note that the general procedure I outline in [Definition 27](#) permits time-varying specifications for $h(\cdot)$.

To predict one-step-ahead GDP inflation, [Krüger et al. \(2017\)](#) again adopt Model 3.10. Concretely, y_t now denotes GDP inflation for quarter t , X_t consists of the additional additional monthly predictors for GDP inflation (observed in the first month of quarter t) listed in Table 3.1 plus a constant. While [Krüger et al. \(2017\)](#)'s description seems to omit the lag of GDP inflation in that variable's nowcasting model, I also include it to have symmetry to the GDP growth case and because it does make a difference; see [Appendix C.5.3](#). As a result, the prior can also be set exactly as described for GDP growth.

Next, the nowcasts for the unemployment rate and the T-Bill are constructed from monthly BVARs. To predict the unemployment rate in quarter t , a monthly VAR(3) is estimated consisting of the monthly unemployment rate plus the additional monthly series listed in Table 3.1, using observations through the first month of quarter t . This auxiliary model is used to generate the remaining two months of quarter t , and the quarterly forecast is formed as the average of the single month of data and the forecasts for the two additional months in the quarter. In the original paper of [Krüger et al. \(2017\)](#), the authors estimate BVARs with stochastic volatility; however, I will estimate a BVAR with fixed coefficients and fixed error variance for simplicity. The particular specification I adopt is a BVAR with a non-Kronecker prior as in [Definition 24](#) (to allow cross-variable shrinkage). To complete the prior specification, I assume the prior satisfies Assumptions 14, 15, and 17. The nowcast for the T-Bill rate is constructed analogously.

Section 3.3 Procedures Given the core BVAR and the additional predictors listed in Table 3.1, one can define an implicit extended VARX with additional regressors, and thus approach the problem of forecast updating using the procedures outlined in [Section 3.3](#).

When the core BVAR uses the first prior (with Kronecker structure), construct the prior for the extended VARX $p(B, \Sigma_B)$ as in [Definition 23](#), reusing Assumptions 14, 16, and 17 to set the values. With this extended prior $p(B, \Sigma_B)$, one can infer $B_{0,1}, B_{0,2}, \Phi_{0,B}$, which are sufficient ingredients to adjust the forecasts according to the procedure of [Definition 26](#).

When the core BVAR uses the second prior (without Kronecker structure), construct the prior for the extended VARX $p(B, \Sigma_B)$ as in [Definition 24](#), but now under Assumptions 14, 15, and 17. This gives sufficient ingredients to adjust according to [Definition 28](#).

Remark. (Sparsity in Nowcasting versus my Procedure) My procedure adjusts the forecast of every outcome

using all available information listed in Table 3.1. In contrast, the four nowcasts are constructed from four separate auxiliary models that exploit a subset of available information. This effectively imposes sparsity in predicting the outcomes, which may not be appropriate for macroeconomic forecasting (Giannone et al., 2021). It is well known that many economic time series exhibit strong comovement (Stock and Watson, 2002). My recommended procedure instead jointly extracts the marginal information and signal from the full dataset. And by following the recommended procedure, shrinkage is implicitly imposed to penalize overfitting.

3.4.3 Accuracy Comparison

In this section, I compare accuracy of the forecast distributions from (i) the core BVAR (no conditioning or tilting), (ii) fully tilting towards nowcasts, and (iii) tilting as recommended by this paper. I evaluate on the sample period defined in Section 3.4.1, while Appendix C.5.3 contains results for different sample starting periods and subsamples—none of which differ qualitatively from the results of this section or suggest that one method uniformly dominates. As the main evaluation metrics, I consider root mean squared error (RMSE) as well as empirical coverage probabilities for the forecast distributions, the latter of which is one advantage of tilting as opposed to conditioning (which collapse the distribution).

Exercise Description Figure 3.4 gives a visual description of the exercise. The black line represents the real-time history of the outcome, i.e. the data as it would have existed at the time and used in estimation. Green shading represents the multi-step joint forecast distribution of (1), constructed from draws taken after initial estimation of the model. Blue and red represent the tilted forecast distributions for (2) and (3), whose entire forecast distribution was adjusted based on the external information and updated forecast for the first period. The gray line represents the realized values. For GDP growth and inflation, which are subject to revision, the realized values are the second release, following Romer and Romer (2000) and Krüger et al. (2017).

Before stating results, I state explicitly the tilting constraints imposed to adjust the forecast distribution from (1) to (2) and (3), using the notation of Section 3.2.3. The unconditional core BVAR (1) provides draws $\{y_{T+1:T+4}^{(d)}\}_{d=1}^D$ from the posterior predictive (joint forecast) distribution, $f_A(y_{T+1:T+4}|Y)$. Given the four nowcasts for the elements of y_{T+1} , all stacked into $\bar{g}_{Now}$, I formulate a constraint

$$\bar{g}_{Now} = g(y_{T+1:T+h}) = y_{T+1}. \quad (3.11)$$

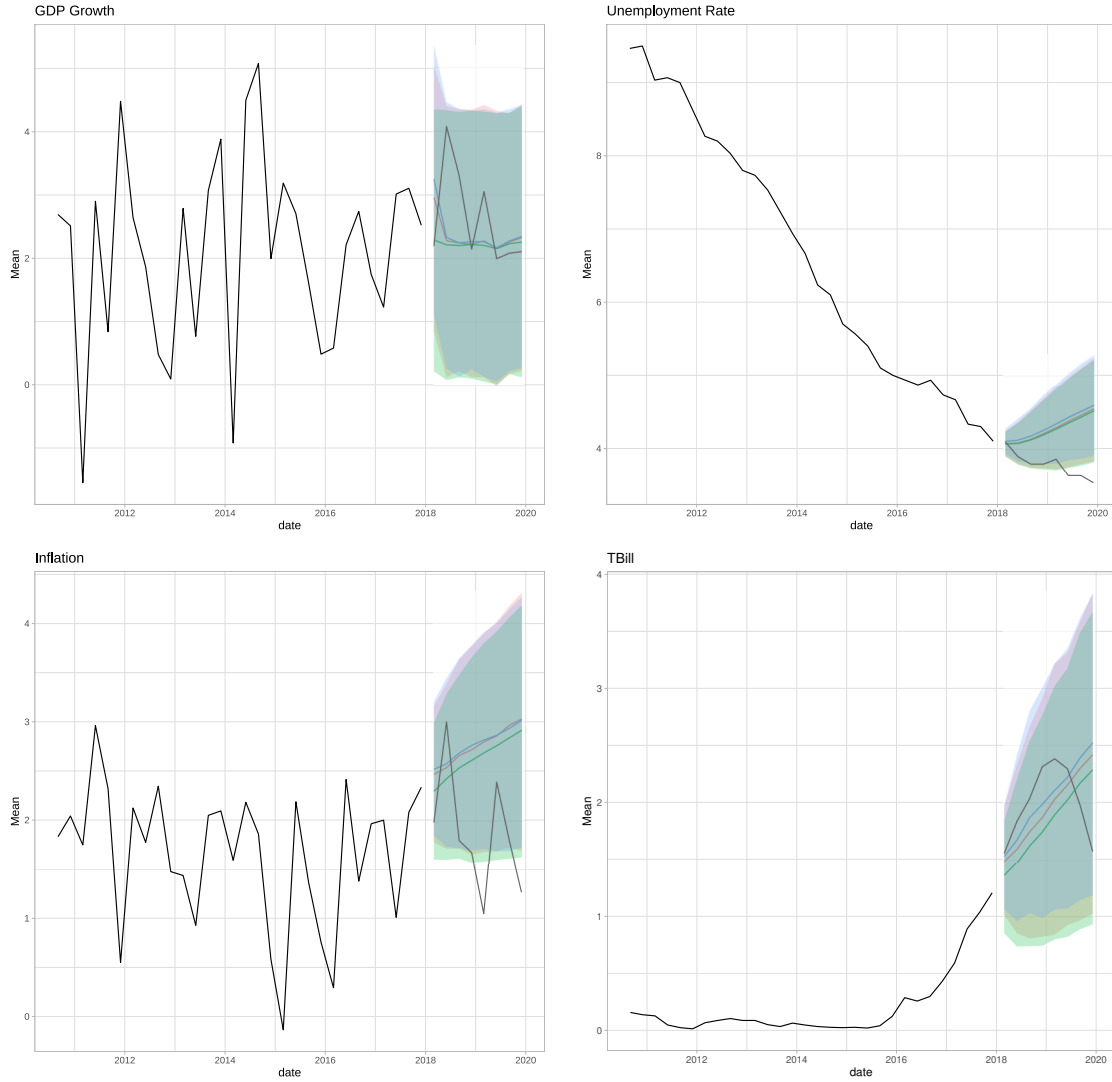


Figure 3.4: Comparison of forecast distributions. Black is the historical series in real-time, as it would have existed when standing in the first forecast period. Green is the forecast distribution from the core BVAR. Blue represents the distribution tilted based on nowcasts, while red represents the distribution tilted based on my recommended procedure. Finally, the gray line is the realized data. Following [Romer and Romer \(2000\)](#), I depict the second release for each forecast period (so that the gray line is an amalgam of vintages). Subtracting any given median forecast line from the gray line gives the errors used to compute the RMSEs.

Using this constraint and the draws $\{y_{T+1:T+4}^{(d)}\}_{d=1}^D$ from (1), I solve the discrete version of the tilting problem in [Definition 25](#) and find updated probabilities/weights p_d^* such that the constraint holds. In [Figure 3.4](#), the blue median line and quantiles are computed using these weights and the original draws. In this way, the adjustment in the first period also propagates and adjusts the full path. To get the adjusted distribution of my procedure (3, in red), I apply the same approach, but using $\bar{g}_{Adj}$ in the constraint, which instead contains the adjusted one-step-ahead forecasts that follow from my procedure.

Baseline Evaluation under Prior 1 [Table 3.2](#) states the RMSE for the competing forecasts, over the full sample. For the BVAR and baseline forecasts, I consider the first specification described in [Section 3.4.2](#), which is the prior with Kronecker structure in the prior variance for coefficients.

Lastly, note that there are two periods where the tilting procedure cannot be implemented because the constraint for the first period forecast lies in the tails of the joint forecast distribution. In other words, the finite number of draws from the core BVAR cannot be reweighted to satisfy the constraint. This is most common when there is a pronounced economic shock. In these cases, it is still possible to implement entropic tilting by appealing to the analytical tilting problem (rather than discrete problem and reweighting), mentioned in [Section 3.2.3](#) and outlined in the appendix of [Antolín-Díaz et al. \(2021\)](#). However, such large forecast revisions that move into the tail of the joint forecast distribution often signify that tilting may not be appropriate. For that reason, and for simplicity, I drop those dates when computing the RMSE in the sample.¹⁹

The main takeaway from [Table 3.2](#) is that the recommended procedure compares favorably and essentially performs as well as tilting towards nowcasts. The improvements from additional information (nowcast or my procedure) are overall quite modest, but the recommended procedure does appear to exploit the information as efficiently as the nowcasting approach in [Krüger et al. \(2017\)](#). The largest exception is the T-Bill. This is perhaps not surprising because the T-Bill is a highly persistent series that is genuinely observed daily with perfect precision. In contrast, GDP growth is not perfectly observed daily, is only available at quarterly frequency, and is quite noisy. Therefore, the nowcasting problem for GDP is similar to a signal extraction problem where one must extract information from related but distinct noisy data series. Note also that the improvement for the T-Bill nearly dissipates beyond the first quarter. Lastly, failing to include the dummies in the regression procedure does slightly reduce accuracy. This suggests that implicit penalization imposed by including dummies may help accuracy somewhat, although the effect is quite modest for this particular

¹⁹For [Table 3.2](#), the dropped dates are Those dates are 2008:Q4 and 2009:Q3. Unsurprisingly, these are dates correspond to time periods where there are recessions or sharp changes in series.

Horizon	Model	GDP Growth	Inflation	Unemployment	T-Bill
h = 1	(1) BVAR	2.4	1.00	0.24	0.36
	(2) Nowcasts	1.8	1.00	0.13	0.06
	(3) Procedure	1.7	0.82	0.12	0.11
	(4) No Dummies	1.8	0.92	0.14	0.22
h = 2	(1) BVAR	2.7	1.10	0.42	0.66
	(2) Nowcasts	2.6	1.10	0.33	0.40
	(3) Procedure	2.6	1.00	0.31	0.43
	(4) No Dummies	2.6	1.10	0.32	0.50
h = 3	(1) BVAR	2.8	1.20	0.67	0.93
	(2) Nowcasts	2.7	1.20	0.54	0.70
	(3) Procedure	2.8	1.10	0.53	0.71
	(4) No Dummies	2.8	1.20	0.55	0.76
h = 4	(1) BVAR	2.8	1.40	0.89	1.20
	(2) Nowcasts	2.8	1.30	0.77	0.97
	(3) Procedure	2.8	1.30	0.77	0.98
	(4) No Dummies	2.8	1.30	0.79	1.00

Table 3.2: Comparison of RMSE for the different outcomes, where the Core BVAR is estimated under Prior 1. The BVAR (1) represents the unconditional forecasts under the first prior, as described in [Section 3.4.2](#). The accuracy for (2) and (3) represents the accuracy of the tilted forecast distributions. The accuracy of (4) is from the procedure, dropping dummies and hence penalization. The evaluation period runs from 1992:Q1-2019:Q4, dropping the dates 2008:Q4 & 2009:Q3 where tilting cannot be implemented.

prior and application.

As another evaluation metric, [Table 3.3](#) reports coverage statistics for the forecast distribution. I consider coverage of the 25th-75th percentile interval as well as the 5th-95th percentile interval. The baseline BVAR already exhibits overcoverage, likely because the estimation period includes data from before the Great Moderation i.e. (1970-1984). Consistent with the RMSE results, coverage increases.

Remark. (Interpretation of Nowcast and versus My Procedure) It is worth noting that any forecast improvement in the nowcast approach stems from both additional information *and* specification changes. By fully tilting towards the one-step-ahead nowcast (as in [Krüger et al. \(2017\)](#)), the approach discards the one-step-ahead forecasts from the core BVAR and, via tilting, simply replaces these one-step-ahead forecast with forecasts from these alternative models. As a result, the core BVAR serves only to extrapolate the one-step-ahead nowcast to longer horizons. Instead, my procedure has the interpretation of expanding and mimicking estimation of a nesting extended model.

To more fully explain the distinction, consider the nowcast of GDP growth. The nowcast regresses quarterly GDP growth on a single lag of quarterly GDP growth and two additional monthly time series. In the

Horizon	Model	GDP Growth		Inflation		Unemployment		T-Bill	
		25-75	5-95	25-75	5-95	25-75	5-95	25-75	5-95
h = 1	(1) BVAR	0.64	0.97	0.56	0.95	0.62	0.93	0.89	1.00
	(2) Nowcasts	0.82	0.97	0.57	0.95	0.81	0.99	1.00	1.00
	(3) Procedure	0.83	1.00	0.64	0.96	0.84	0.99	1.00	1.00
	(4) No Dummies	0.76	0.98	0.67	0.95	0.79	1.00	0.98	1.00
h = 2	(1) BVAR	0.66	0.96	0.58	0.98	0.64	0.92	0.84	0.98
	(2) Nowcasts	0.68	0.97	0.58	0.98	0.72	0.96	0.93	1.00
	(3) Procedure	0.69	0.97	0.64	0.98	0.76	0.96	0.93	1.00
	(4) No Dummies	0.66	0.96	0.59	0.97	0.69	0.96	0.91	1.00
h = 3	(1) BVAR	0.65	0.93	0.62	0.98	0.63	0.90	0.79	0.96
	(2) Nowcasts	0.67	0.94	0.65	0.98	0.74	0.95	0.89	1.00
	(3) Procedure	0.64	0.94	0.63	0.99	0.76	0.95	0.88	0.99
	(4) No Dummies	0.68	0.94	0.64	0.99	0.70	0.95	0.85	0.99
h = 4	(1) BVAR	0.64	0.96	0.58	0.99	0.58	0.90	0.79	0.95
	(2) Nowcasts	0.64	0.95	0.60	0.99	0.68	0.93	0.88	0.98
	(3) Procedure	0.65	0.95	0.63	1.00	0.71	0.93	0.87	0.97
	(4) No Dummies	0.65	0.95	0.63	1.00	0.63	0.92	0.83	0.97

Table 3.3: Comparison of coverage for the different outcomes, where the Core BVAR is estimated under Prior 1. The BVAR (1) represents the unconditional forecasts under the first prior, as described in [Section 3.4.2](#). The accuracy for (2) and (3) represents the accuracy of the tilted forecast distributions. The accuracy of (4) is from the procedure, dropping dummies and hence penalization. The evaluation period runs from 1992:Q1-2019:Q4, dropping the dates 2008:Q4 & 2009:Q3 where tilting cannot be implemented.

Horizon	Model	GDP Growth	Inflation	Unemployment	T-Bill
h = 1	(1) BVAR	2.2	1.00	0.24	0.39
	(2) Nowcasts	1.8	1.00	0.13	0.06
	(3) Procedure	1.9	0.93	0.13	0.15
	(4) No Dummies	1.7	0.91	0.13	0.16
h = 2	(1) BVAR	2.4	1.10	0.41	0.72
	(2) Nowcasts	2.4	1.10	0.33	0.41
	(3) Procedure	2.5	1.10	0.34	0.52
	(4) No Dummies	1.4	1.00	0.31	0.50
h = 3	(1) BVAR	2.5	1.20	0.63	1.00
	(2) Nowcasts	2.5	1.20	0.53	0.75
	(3) Procedure	2.5	1.20	0.54	0.84
	(4) No Dummies	2.5	1.20	0.51	0.81
h = 4	(1) BVAR	2.4	1.30	0.83	1.30
	(2) Nowcasts	2.5	1.30	0.74	1.00
	(3) Procedure	2.5	1.30	0.75	1.10
	(4) No Dummies	2.5	1.30	0.72	1.10

Table 3.4: Comparison of RMSE for the different outcomes, where the Core BVAR is estimated under Prior 2. The BVAR (1) represents the unconditional forecasts under the first prior, as described in [Section 3.4.2](#). The accuracy for (2) and (3) represents the accuracy of the tilted forecast distributions. The accuracy of (4) is from the procedure, dropping dummies and hence penalization. The evaluation period runs from 1992:Q1-2019:Q4. I drop 2008:Q4, where tilting cannot be implemented.

original core BVAR, GDP growth is *already* regressed on p lags of GDP growth, as well as additional quarterly time series. Therefore, it is not clear whether the improvement in performance comes from the marginal information in the additional regressors or the change to a simpler single lag model that also excludes non-GDP-related macro signals like quarterly inflation, unemployment, and interest rates. Therefore, even if one ultimately is interested in fully tilting towards a nowcast, my procedure can serve as a useful supplementary decomposition to more cleanly characterize the source of improvements: marginal information versus change in model specification.

Evaluation under Prior 2 Next, Table [3.4](#) states the RMSE for the competing forecasts when the core BVAR is estimated under the second specification described in [Section 3.4.2](#), while Table [3.5](#) reports coverage.²⁰ Recall that this is the prior *without* Kronecker structure in the prior variance for coefficients. This alternative comparison is useful both intrinsically (to relax restrict Kronecker structure²¹) and as a guide for applying the procedure when exact equivalence does not hold.

²⁰The only time period where reweighting is not possible is 2008:Q4.

²¹See discussion in [Section 3.2.2](#).

Horizon	Model	GDP Growth		Inflation		Unemployment		T-Bill	
		25-75	5-95	25-75	5-95	25-75	5-95	25-75	5-95
h = 1	(1) BVAR	0.72	0.98	0.56	0.93	0.68	0.93	0.91	0.99
	(2) Nowcasts	0.83	0.98	0.53	0.93	0.83	1.00	1.00	1.00
	(3) Procedure	0.80	0.98	0.60	0.95	0.84	1.00	0.99	1.00
	(3) No Dummies	0.86	0.99	0.65	0.95	0.83	0.99	1.00	1.00
h = 2	(1) BVAR	0.71	0.97	0.51	0.98	0.70	0.93	0.82	0.96
	(2) Nowcasts	0.74	0.97	0.56	0.98	0.72	0.96	0.94	1.00
	(3) Procedure	0.69	0.98	0.55	0.98	0.68	0.96	0.91	1.00
	(3) No Dummies	0.72	0.98	0.58	0.97	0.74	0.96	0.92	1.00
h = 3	(1) BVAR	0.73	0.97	0.61	0.98	0.71	0.91	0.78	0.95
	(2) Nowcasts	0.73	0.96	0.62	0.98	0.68	0.95	0.88	0.97
	(3) Procedure	0.70	0.96	0.62	0.97	0.72	0.93	0.87	0.96
	(3) No Dummies	0.72	0.96	0.61	0.98	0.73	0.96	0.85	0.98
h = 4	(1) BVAR	0.73	0.97	0.57	0.97	0.60	0.92	0.76	0.93
	(2) Nowcasts	0.69	0.97	0.56	0.97	0.67	0.93	0.81	0.96
	(3) Procedure	0.67	0.97	0.57	0.98	0.68	0.93	0.80	0.94
	(3) No Dummies	0.71	0.97	0.59	0.98	0.67	0.93	0.80	0.96

Table 3.5: Comparison of coverage for the different outcomes, where the Core BVAR is estimated under Prior 2. The BVAR (1) represents the unconditional forecasts under the second prior, as described in [Section 3.4.2](#). The coverage for (2) and (3) represents the coverage of the tilted forecast distributions. The evaluation period runs from 1992:Q1-2019:Q4, dropping the date 2008:Q4 where tilting cannot be implemented.

As the table shows, even though exact re-estimation equivalence does not hold, my recommended procedure improves upon forecast accuracy relative to the core BVAR and again compares favorably to tilting on nowcasts. As can be seen from the table, for one-step-ahead forecasts, nowcast accuracy remains the same, since I fully tilt towards the nowcasts, which implies the core BVAR specification is irrelevant. Performance is somewhat worse for GDP, no more than 10 basis points. The procedure mostly improves upon forecasts of the core BVAR wherever nowcast does. The only case where there is a notable difference is, again, for the T-Bill. Relative to the Kronecker case, the exclusion of dummies does improve performance somewhat for GDP, although the differences are modest.

Baseline Evaluation under Random Forest Adjustment Finally, I do a brief comparison between nowcasting, my procedure, and adjustment using a more flexible machine learning model for the errors, following [Definition 27](#). For the core BVAR, I adopt Prior 1, and so the relevant comparison table is [Table 3.2](#); the accuracy for the BVAR, nowcasts, and regression-based adjustments match by construction.

For the model for $h(\cdot)$, I use the `randomForest` package in R ([Liaw and Wiener, 2002](#)) to estimate

Horizon	Model	GDP Growth	Inflation	Unemployment	T-Bill
h = 1	(1) BVAR	2.4	1.00	0.24	0.360
	(2) Nowcasts	1.8	1.00	0.13	0.061
	(3) Procedures				
	(a) Regression	1.7	0.82	0.12	0.110
	(b) Random Forest	2.0	0.91	0.21	0.280
h = 2	(1) BVAR	2.7	1.10	0.42	0.660
	(2) Nowcasts	2.5	1.10	0.33	0.390
	(3) Procedures				
	(a) Regression	2.5	1.00	0.31	0.430
	(b) Random Forest	2.6	1.00	0.38	0.570
h = 3	(1) BVAR	2.8	1.20	0.67	0.930
	(2) Nowcasts	2.7	1.20	0.53	0.700
	(3) Procedures				
	(a) Regression	2.7	1.20	0.52	0.710
	(b) Random Forest	2.8	1.20	0.61	0.820
h = 4	(1) BVAR	2.8	1.40	0.89	1.200
	(2) Nowcasts	2.8	1.40	0.76	0.990
	(3) Procedures				
	(a) Regression	2.8	1.30	0.76	0.980
	(b) Random Forest	2.8	1.30	0.84	1.100

Table 3.6: Comparison of RMSE for the different outcomes, where the Core BVAR is estimated under Prior 1. The BVAR (1) represents the unconditional forecasts under the first prior, as described in [Section 3.4.2](#). The accuracy for (2) and (3) is the accuracy of the adjustment procedures. The evaluation period runs from 1993:Q1-2019:Q4. I drop the same periods as for [Table 3.2](#) to maintain comparability.

a function $h(\cdot)$ under default options for the package—rather than regression—and I follow exactly the procedure of [Definition 27](#) to adjust to the forecast distribution. This mimics using an off-the-shelf machine learning approach to try to improve upon forecast errors, similar to the logic of boosting, which is familiar from the machine learning literature.

[Table 3.6](#) shows that use of a more general model in this case does lead to an improvement over baseline BVAR forecasts; however, the gains are not as good as the regression-based adjustment. Overall, the results do not suggest large gains in the current context, but I view this as an area of future research. Even though I used a more flexible model for $h(\cdot)$ in [Definition 27](#), I still did not expand the set of predictors. These results suggest that real gains would likely need to come from the use of more data. Moreover, tuning and exploration of alternative machine learning algorithms in the adjustment step might yield improvements beyond what I was able to realize.

3.5 Conclusion

Above I have shown how one can define a tilting procedure for BVAR forecast distributions that is exactly equivalent (in mean) or closely related to re-estimation of an expanded model. In a representative example

from the literature, I showed that this procedure compares favorably to other tilting procedures that define and estimate separate auxiliary models outcome-by-outcome, based on accuracy. Instead, it is possible to improve the forecast distribution by exploiting a cross-section of additional predictor variables in a single post-estimation step—avoiding overfitting and reusing output from originally estimating a baseline core BVAR(X)—by appealing to an implicit expanded model. In this way, entropic tilting can be motivated not as an ad hoc adjustment, but as a procedure closely related to the ideal of re-estimation. Future work could build upon these results to exploit even larger cross-sections of the data.

Bibliography

- Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge (2017) “When should you adjust standard errors for clustering?” Technical report, National Bureau of Economic Research.
- Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge (2020) “Sampling-Based versus Design-Based Uncertainty in Regression Analysis,” *Econometrica*, Vol. 88, pp. 265–296.
- Altonji, Joseph G. and Lewis M. Segal (1996) “Small-Sample Bias in GMM Estimation of Covariance Structures,” *Journal of Business & Economic Statistics*, Vol. 14, pp. 353–366, DOI: <http://dx.doi.org/10.1080/07350015.1996.10524661>.
- Alvarez, Fernando, Hervé Le Bihan, and Francesco Lippi (2016) “The Real Effects of Monetary Shocks in Sticky Price Models: A Sufficient Statistic Approach,” *American Economic Review*, Vol. 106, pp. 2817–2851, URL: <https://www.aeaweb.org/articles?id=10.1257/aer.20140500>, DOI: <http://dx.doi.org/10.1257/aer.20140500>.
- Alvarez, Fernando and Francesco Lippi (2014) “Price Setting With Menu Cost for Multiproduct Firms,” *Econometrica*, Vol. 82, pp. 89–135, URL: <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA10662>, DOI: <http://dx.doi.org/https://doi.org/10.3982/ECTA10662>.
- Andrews, Donald WK (1991) “Heteroskedasticity and autocorrelation consistent covariance matrix estimation,” *Econometrica: Journal of the Econometric Society*, pp. 817–858.
- Antolín-Díaz, Juan, Ivan Petrella, and Juan F. Rubio-Ramírez (2021) “Structural scenario analysis with SVARs,” *Journal of Monetary Economics*, Vol. 117, pp. 798–815, URL: <https://www.sciencedirect.com/science/article/pii/S0304393220300751>, DOI: <http://dx.doi.org/https://doi.org/10.1016/j.jmoneco.2020.06.001>.

- Aronow, Peter M, Cyrus Samii et al. (2017) “Estimating average causal effects under general interference, with application to a social network experiment,” *The Annals of Applied Statistics*, Vol. 11, pp. 1912–1947.
- Auclert, Adrien, Bence Bardóczy, Matthew Rognlie, and Ludwig Straub (2021) “Using the Sequence-Space Jacobian to Solve and Estimate Heterogeneous-Agent Models,” *Econometrica*, Vol. 89, pp. 2375–2408.
- Bañbura, Marta, Domenico Giannone, Michele Modugno, and Lucrezia Reichlin (2013) “Now-casting and the real-time data flow,” in *Handbook of economic forecasting*, Vol. 2: Elsevier, pp. 195–237.
- Banerjee, S, AE Gelfand, John R Knight, and CF Sirmans (2004) “Spatial modeling of house prices using normalized distance-weighted sums of stationary processes,” *Journal of Business & Economic Statistics*, Vol. 22, pp. 206–213.
- Barrios, Thomas, Rebecca Diamond, Guido W. Imbens, and Michal Kolesár (2012) “Clustering, Spatial Correlations, and Randomization Inference,” *Journal of the American Statistical Association*, Vol. 107, pp. 578–591, DOI: <http://dx.doi.org/10.1080/01621459.2012.682>.
- Baum-Snow, Nathaniel and Fernando Ferreira (2015) “Causal Inference in Urban and Regional Economics,” in *Handbook of Regional and Urban Economics*, Vol. 5: Elsevier, pp. 3–68.
- Baum-Snow, Nathaniel and Justin Marion (2009) “The effects of low income housing tax credit developments on neighborhoods,” *Journal of Public Economics*, Vol. 93, pp. 654–666.
- Bañbura, Marta, Federica Brenna, Joan Paredes, and Francesco Ravazzolo (2021) “Combining Bayesian VARs with survey density forecasts: does it pay off?,” Working Paper Series 2543, European Central Bank.
- Bera, Anil K and Pradosh Simlai (2005) “Testing spatial autoregressive model and a formulation of spatial ARCH (SARCH) model with applications.”
- Blattman, Christopher, Donald Green, Daniel Ortega, and Santiago Tobón (2017) “Place-based interventions at scale: The direct and spillover effects of policing and city services on crime,” Technical report, National Bureau of Economic Research.
- Bolthausen, E. (1982) “On the Central Limit Theorem for Stationary Mixing Random Fields,” *The Annals of Probability*, Vol. 10, pp. 1047 – 1050, URL: <https://doi.org/10.1214/aop/1176993726>, DOI: <http://dx.doi.org/10.1214/aop/1176993726>.

- Borusyak, Kirill and Peter Hull (2023) “Nonrandom Exposure to Exogenous Shocks,” *Econometrica*, Vol. 91, pp. 2155–2185, URL: <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA19367>, DOI: <http://dx.doi.org/https://doi.org/10.3982/ECTA19367>.
- Brown, Jessica H (2018) “Does public pre-k have unintended consequences on the child care market for infants and toddlers?” *Princeton University Industrial Relations Section Working Paper*, Vol. 626.
- Cameron, A. Colin and Douglas L. Miller (2015) “A Practitioner’s Guide to Cluster-Robust Inference,” *Journal of Human Resources*, Vol. 50, pp. 317–372, URL: <https://jhr.uwpress.org/content/50/2/317>, DOI: <http://dx.doi.org/10.3368/jhr.50.2.317>.
- Carriero, Andrea, Todd E. Clark, and Massimiliano Marcellino (2015a) “Bayesian VARs: Specification Choices and Forecast Accuracy,” *Journal of Applied Econometrics*, Vol. 30, pp. 46–73, URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2315>, DOI: <http://dx.doi.org/https://doi.org/10.1002/jae.2315>.
- (2015b) “Realtime nowcasting with a Bayesian mixed frequency model with stochastic volatility,” *Journal of the Royal Statistical Society Series A*, Vol. 178, pp. 837–862, URL: <https://ideas.repec.org/a/bla/jorssa/v178y2015i4p837-862.html>.
- Carriero, Andrea, Todd E Clark, Massimiliano Marcellino, and Elmar Mertens (2022) “Addressing COVID-19 outliers in BVARs with stochastic volatility,” *Review of Economics and Statistics*, pp. 1–38.
- Chang, Minsu, Xiaohong Chen, and Frank Schorfheide (2023) “Heterogeneity and Aggregate Fluctuations,” Manuscript, University of Pennsylvania.
- Chen, Louis H. Y. and Qi-Man Shao (2004) “Normal approximation under local dependence,” *Ann. Probab.*, Vol. 32, pp. 1985–2028, URL: <https://doi.org/10.1214/009117904000000450>, DOI: <http://dx.doi.org/10.1214/009117904000000450>.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018) “Double/debiased machine learning for treatment and structural parameters,” *The Econometrics Journal*, Vol. 21, pp. C1–C68, URL: <https://doi.org/10.1111/ectj.12097>, DOI: <http://dx.doi.org/10.1111/ectj.12097>.
- Chib, Siddhartha and Edward Greenberg (1994) “Bayes inference in regression models with ARMA (p, q)

- errors,” *Journal of Econometrics*, Vol. 64, pp. 183–206, URL: <https://EconPapers.repec.org/RePEc:eee:econom:v:64:y:1994:i:1-2:p:183-206>.
- Cogley, Timothy, Sergei Morozov, and Thomas J. Sargent (2005) “Bayesian fan charts for U.K. inflation: Forecasting and sources of uncertainty in an evolving monetary system,” *Journal of Economic Dynamics and Control*, Vol. 29, pp. 1893–1925, URL: <https://www.sciencedirect.com/science/article/pii/S0165188905001090>, DOI: <http://dx.doi.org/https://doi.org/10.1016/j.jedc.2005.06.005>, Expectations, learning, and monetary policy.
- Conley, Timothy G (1999) “GMM Estimation with Cross-Sectional Dependence,” *Journal of Econometrics*, Vol. 92, pp. 1–45.
- Couture, Victor, Cecile Gaubert, Jessie Handbury, and Erik Hurst (2019) “Income growth and the distributional effects of urban spatial sorting,” Technical report, National Bureau of Economic Research.
- Crump, Richard K., Stefano Eusepi, Domenico Giannone, Eric Qian, and Argia M. Sbordone (2021) “A Large Bayesian VAR of the United States Economy,” Staff Reports 976, Federal Reserve Bank of New York.
- Davis, Lucas W (2011) “The effect of power plants on local housing values and rents,” *Review of Economics and Statistics*, Vol. 93, pp. 1391–1402.
- DeJong, D. N. and C. Dave (2011) *Structural Macroeconometrics*: Princeton University Press, 2nd edition.
- Del Negro, Marco and Frank Schorfheide (2004) “Priors from General Equilibrium Models for Vars,” *International Economic Review*, Vol. 45, pp. 643–673, URL: <http://www.jstor.org/stable/3663534>.
- Diamond, Rebecca and Tim McQuade (2019) “Who wants affordable housing in their backyard? An equilibrium analysis of low-income property development,” *Journal of Political Economy*, Vol. 127, pp. 1063–1117.
- Dickey, James M. (1967) “Matricvariate Generalizations of the Multivariate t Distribution and the Inverted Multivariate t Distribution,” *The Annals of Mathematical Statistics*, Vol. 38, pp. 511–518, URL: <http://www.jstor.org/stable/2239163>.
- Diggle, Peter J (2013) *Statistical analysis of spatial and spatio-temporal point patterns*: CRC press.
- Doan, Thomas, Robert Litterman, and Christopher Sims (1984) “Forecasting and conditional projection using realistic prior distributions,” *Econometric Reviews*, Vol. 3, pp. 1–100, URL: <https://doi.org/10.1080/07474938408800053>, DOI: <http://dx.doi.org/10.1080/07474938408800053>.

- Doh, Taeyoung (2017) “Trend and Uncertainty in the Long-Term Real Interest Rate: Bayesian Exponential Tilting with Survey Data,” Research Working Paper RWP 17-8, Federal Reserve Bank of Kansas City.
- Driscoll, John C. and Aart C. Kraay (1998) “Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data,” *Review of Economics and Statistics*, Vol. 80, pp. 549–560, URL: <https://doi.org/10.1162/003465398557825>, DOI: <http://dx.doi.org/10.1162/003465398557825>.
- Ellen, Ingrid Gould, Johanna Lacoe, and Claudia Ayanna Sharygin (2013) “Do foreclosures cause crime?” *Journal of Urban Economics*, Vol. 74, pp. 59–70.
- Evans, George W. and Seppo Honkapohja (2005) “An Interview with Thomas J. Sargent,” *Macroeconomic Dynamics*, Vol. 9, p. 561–583, DOI: <http://dx.doi.org/10.1017/S1365100505050042>.
- Freedman, Matthew and Annemarie Kuhns (2018) “Supply-side subsidies to improve food access and dietary outcomes: Evidence from the New Markets Tax Credit,” *Urban Studies*, Vol. 55, pp. 3234–3251, URL: <https://doi.org/10.1177/0042098017740285>, DOI: <http://dx.doi.org/10.1177/0042098017740285>.
- Galvão, Ana Beatriz, Anthony Garratt, and James Mitchell (2021) “Does judgment improve macroeconomic density forecasts?,” *International Journal of Forecasting*, Vol. 37, pp. 1247–1260, URL: <https://ideas.repec.org/a/eee/intfor/v37y2021i3p1247-1260.html>, DOI: <http://dx.doi.org/10.1016/j.ijforecast.2021>.
- Georgescu, Dan I., Nicholas J. Higham, and Gareth W. Peters (2018) “Explicit solutions to correlation matrix completion problems, with an application to risk management and insurance,” *Royal Society Open Science*, Vol. 5, pp. 1–11.
- Giacomini, Raffaella and Giuseppe Ragusa (2014) “Theory-coherent forecasting,” *Journal of Econometrics*, Vol. 182, pp. 145–155, URL: <https://ideas.repec.org/a/eee/econom/v182y2014i1p145-155.html>, DOI: <http://dx.doi.org/10.1016/j.jeconom.2014.04>.
- Giannone, Domenico, Michele Lenza, and Giorgio Primiceri (2015) “Prior Selection for Vector Autoregressions,” *The Review of Economics and Statistics*, Vol. 97, pp. 436–451, URL: <https://EconPapers.repec.org/RePEc:tpr:restat:v:97:y:2015:i:2:p:436-451>.
- Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri (2021) “Economic Predictions With Big Data:

- The Illusion of Sparsity,” *Econometrica*, Vol. 89, pp. 2409–2437, URL: <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA17842>, DOI: <http://dx.doi.org/https://doi.org/10.3982/ECTA17842>.
- Hahn, Jinyong (1998) “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica*, Vol. 66, pp. 315–331, URL: <http://www.jstor.org/stable/2998560>.
- Hahn, Jinyong, Guido Kuersteiner, and Maurizio Mazzocco (2020) “Estimation with Aggregate Shocks,” *Review of Economic Studies*, Vol. 87, pp. 1365–1398.
- (2022a) “Central Limit Theory for Combined Cross Section and Time Series With an Application to Aggregate Productivity Shocks,” *Econometric Theory*, DOI: <http://dx.doi.org/10.1017/S0266466622000391>, Forthcoming.
- (2022b) “Joint Time Series and Cross-Section Limit Theory under Mixingale Assumptions,” *Econometric Theory*, Vol. 38, pp. 942–958.
- Hall, P., C.C. Heyde, Z.W. Birnbaum, and E. Lukacs (2014) *Martingale Limit Theory and Its Application*, Communication and Behavior: Elsevier Science, URL: <https://books.google.com/books?id=gqriBQAAQBAJ>.
- Hamilton, James D. (1994) *Time Series Analysis*: Princeton University Press, URL: <http://www.jstor.org/stable/j.ctv14jx6sm>.
- Hansen, LP and JJ Heckman (1996) “The Empirical Foundations of Calibration,” *Journal of Economic Perspectives*, Vol. 10, pp. 87–104, URL: <http://www.jstor.org/stable/10.2307/2138285>.
- Hirano, Keisuke, Guido W Imbens, and Geert Ridder (2003) “Efficient estimation of average treatment effects using the estimated propensity score,” *Econometrica*, Vol. 71, pp. 1161–1189.
- Hong, Han, Michael P Leung, and Jessie Li (2019) “Inference on finite-population treatment effects under limited overlap,” *The Econometrics Journal*, Vol. 23, pp. 32–47, URL: <https://doi.org/10.1093/ectj/utz017>, DOI: <http://dx.doi.org/10.1093/ectj/utz017>.
- Hong, Han, Aprajit Mahajan, and Denis Nekipelov (2015) “Extremum estimation and numerical derivatives,” *Journal of Econometrics*, Vol. 188, pp. 250–263, URL: <https://linkinghub.elsevier.com/retrieve/pii/S0304407615001207>, DOI: <http://dx.doi.org/10.1016/j.jeconom.2014.05.019>.

- Imbens, Guido W (2002) “Generalized Method of Moments and Empirical Likelihood,” *Journal of Business & Economic Statistics*, Vol. 20, pp. 493–506, URL: <http://www.tandfonline.com/doi/abs/10.1198/073500102288618630>, DOI: <http://dx.doi.org/10.1198/073500102288618630>.
- Imbens, Guido W. and Joshua D. Angrist (1994) “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, Vol. 62, pp. 467–475, URL: <http://www.jstor.org/stable/2951620>.
- Imbens, Guido W. and Tony Lancaster (1994) “Combining Micro and Macro Data in Microeconomic Models,” *Review of Economic Studies*, Vol. 61, pp. 655–680, URL: <http://www.jstor.org/stable/2297913>.
- Imbens, Guido W and Donald B Rubin (2015) *Causal inference in statistics, social, and biomedical sciences*: Cambridge University Press.
- Jenish, Nazgul and Ingmar R. Prucha (2009) “Central limit theorems and uniform laws of large numbers for arrays of random fields,” *Journal of Econometrics*, Vol. 150, pp. 86 – 98, URL: <http://www.sciencedirect.com/science/article/pii/S0304407609000475>, DOI: <http://dx.doi.org/https://doi.org/10.1016/j.jeconom.2009.02.009>.
- Kadiyala, K. Rao and Sune Karlsson (1997) “Numerical Methods for Estimation and Inference in Bayesian VAR-Models,” *Journal of Applied Econometrics*, Vol. 12, pp. 99–132, URL: <http://www.jstor.org/stable/2284907>.
- Karlsson, Sune (2013) “Chapter 15 - Forecasting with Bayesian Vector Autoregression,” in Graham Elliott and Allan Timmermann eds. *Handbook of Economic Forecasting*, Vol. 2 of Handbook of Economic Forecasting: Elsevier, pp. 791–897, URL: <https://www.sciencedirect.com/science/article/pii/B9780444627315000154>, DOI: <http://dx.doi.org/https://doi.org/10.1016/B978-0-444-62731-5.00015-4>.
- Knotek, Edward S. and Saeed Zaman (2019) “Financial nowcasts and their usefulness in macroeconomic forecasting,” *International Journal of Forecasting*, Vol. 35, pp. 1708–1724, URL: <https://ideas.repec.org/a/eee/intfor/v35y2019i4p1708-1724.html>, DOI: <http://dx.doi.org/10.1016/j.ijforecast.2018>.
- Koenker, Roger and Gilbert Bassett (1978) “Regression Quantiles,” *Econometrica*, Vol. 46, pp. 33–50.
- Koop, Gary, Stuart McIntyre, and James Mitchell (2020) “UK regional nowcasting using a mixed frequency vector auto-regressive model with entropic tilting,” *Journal of the Royal Statistical Society Series A*,

- Vol. 183, pp. 91–119, URL: <https://ideas.repec.org/a/bla/jorssa/v183y2020i1p91-119.html>, DOI: <http://dx.doi.org/10.1111/rssa.12491>.
- Krüger, Fabian, Todd E. Clark, and Francesco Ravazzolo (2017) “Using Entropic Tilting to Combine BVAR Forecasts With External Nowcasts,” *Journal of Business & Economic Statistics*, Vol. 35, pp. 470–485, URL: <https://doi.org/10.1080/07350015.2015.1087856>, DOI: <http://dx.doi.org/10.1080/07350015.2015.1087856>.
- Kydland, Finn E and Edward C Prescott (1996) “The Computational Experiment: An Econometric Tool,” *Journal of Economic Perspectives*, Vol. 10, pp. 69–85, URL: <http://www.jstor.org/stable/10.2307/2138284><http://pubs.aeaweb.org/doi/10.1257/jep.10.1.69>, DOI: <http://dx.doi.org/10.1257/jep.10.1.69>.
- Lazarus, Eben, Daniel J. Lewis, James H. Stock, and Mark W. Watson (2018) “HAR Inference: Recommendations for Practice,” *Journal of Business & Economic Statistics*, Vol. 36, pp. 541–574.
- Leung, Michael P. (2022) “Causal Inference Under Approximate Neighborhood Interference,” *Econometrica*, Vol. 90, pp. 267–293, URL: <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA17841>, DOI: <http://dx.doi.org/https://doi.org/10.3982/ECTA17841>.
- Liaw, Andy and Matthew Wiener (2002) “Classification and Regression by randomForest,” *R News*, Vol. 2, pp. 18–22, URL: <https://CRAN.R-project.org/doc/Rnews/>.
- Lindenmeyer, Guilherme Schultz and Hudson da Silva Torrent (2023) “Boosting and Predictability of Macroeconomic Variables: Evidence from Brazil,” *Computational Economics*, URL: <https://api.semanticscholar.org/CorpusID:260922088>.
- Litterman, Robert B (1980) *Bayesian procedure for forecasting with vector autoregressions*: Massachusetts Institute of Technology.
- Liu, Laura and Mikkel Plagborg-Møller (2023) “Full-information estimation of heterogeneous agent models using macro and micro data,” *Quantitative Economics*, Vol. 14, pp. 1–35, URL: <https://www.econometricsociety.org/doi/10.3982/QE1810>, DOI: <http://dx.doi.org/10.3982/QE1810>.
- Luo, Ye and Martin Spindler (2017) “L2-Boosting for Economic Applications,” *The American Economic Review*, Vol. 107, pp. 270–273, URL: <http://www.jstor.org/stable/44250403>.

- Marcellino, Massimiliano, James H. Stock, and Mark W. Watson (2006) “A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series,” *Journal of Econometrics*, Vol. 135, pp. 499–526, URL: <https://www.sciencedirect.com/science/article/pii/S030440760500165X>, DOI: <http://dx.doi.org/https://doi.org/10.1016/j.jeconom.2005.07.020>.
- Marcelo C. Medeiros, Álvaro Veiga, Gabriel F. R. Vasconcelos and Eduardo Zilberman (2021) “Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods,” *Journal of Business & Economic Statistics*, Vol. 39, pp. 98–119, URL: <https://doi.org/10.1080/07350015.2019.1637745>, DOI: <http://dx.doi.org/10.1080/07350015.2019.1637745>.
- Masini, Ricardo P., Marcelo C. Medeiros, and Eduardo F. Mendes (2023) “Machine learning advances for time series forecasting,” *Journal of Economic Surveys*, Vol. 37, pp. 76–111, URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/joes.12429>, DOI: <http://dx.doi.org/https://doi.org/10.1111/joes.12429>.
- Mavroeidis, Sophocles, Mikkel Plagborg-Møller, and James H. Stock (2014) “Empirical Evidence on Inflation Expectations in the New Keynesian Phillips Curve,” *Journal of Economic Literature*, Vol. 52, pp. 124–188, URL: <http://pubs.aeaweb.org/doi/abs/10.1257/jel.52.1.124>, DOI: <http://dx.doi.org/10.1257/jel.52.1.124>.
- McKay, Alisdair, Emi Nakamura, and Jón Steinsson (2016) “The Power of Forward Guidance Revisited,” *American Economic Review*, Vol. 106, pp. 3133–58, URL: <https://www.aeaweb.org/articles?id=10.1257/aer.20150063>, DOI: <http://dx.doi.org/10.1257/aer.20150063>.
- Miguel, Edward and Michael Kremer (2004) “Worms: identifying impacts on education and health in the presence of treatment externalities,” *Econometrica*, Vol. 72, pp. 159–217.
- Miranda-Agrippino, Silvia and Giovanni Ricco (2021) “The Transmission of Monetary Policy Shocks,” *American Economic Journal: Macroeconomics*, Vol. 13, pp. 74–107, URL: <https://www.aeaweb.org/articles?id=10.1257/mac.20180124>, DOI: <http://dx.doi.org/10.1257/mac.20180124>.
- Mobarak, Ahmed Mushfiq and Mark R. Rosenzweig (2013) “Informal Risk Sharing, Index Insurance, and Risk Taking in Developing Countries,” *American Economic Review*, Vol. 103, pp. 375–80, URL: <https://www.aeaweb.org/articles?id=10.1257/aer.103.3.375>, DOI: <http://dx.doi.org/10.1257/aer.103.3.375>.

- Mongey, Simon and Jerome Williams (2017) “Firm dispersion and business cycles: Estimating aggregate shocks using panel data,” Manuscript, New York University.
- Müller, Ulrich K. (2014) “HAC Corrections for Strongly Autocorrelated Time Series,” *Journal of Business & Economic Statistics*, Vol. 32, pp. 311–322, URL: <http://www.jstor.org/stable/43702759>.
- Müller, Ulrich K. and Mark W. Watson (2022) “Spatial Correlation Robust Inference,” *Econometrica*, Vol. 90, pp. 2901–2935, URL: <https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA19465>, DOI: <http://dx.doi.org/https://doi.org/10.3982/ECTA19465>.
- Nakamura, Emi and Jón Steinsson (2018) “Identification in Macroeconomics,” *Journal of Economic Perspectives*, Vol. 32, pp. 59–86.
- Newey, Whitney K and Daniel L McFadden (1994) “Large Sample Estimation and Hypothesis Testing,” in Robert F. Engle and Daniel L. McFadden eds. *Handbook of Econometrics, Volume IV*: Elsevier, Chap. 36, pp. 2111–2245, DOI: [http://dx.doi.org/10.1016/S1573-4412\(05\)80005-4](http://dx.doi.org/10.1016/S1573-4412(05)80005-4).
- Newey, Whitney K and Kenneth D West (1986) “A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix,” Technical report, National Bureau of Economic Research.
- Osiewalski, Jacek (2011) “Bayesian Variations on the Frisch and Waugh Theme,” *Central European Journal of Economic Modelling and Econometrics*, Vol. 3, pp. 39–47, URL: <https://EconPapers.repec.org/RePEc:psc:journl:v:3:y:2011:i:1:p:39-47>.
- Pollmann, Michael (2023) “Causal Inference for Spatial Treatments.”
- Primiceri, Giorgio E. (2005) “Time Varying Structural Vector Autoregressions and Monetary Policy,” *The Review of Economic Studies*, Vol. 72, pp. 821–852, URL: <https://doi.org/10.1111/j.1467-937X.2005.00353.x>, DOI: <http://dx.doi.org/10.1111/j.1467-937X.2005.00353.x>.
- Qian, Jiahe (2020) “Variance Estimation With Complex Data and Finite Population Correction—A Paradigm for Comparing Jackknife and Formula-Based Methods for Variance Estimation,” *ETS Research Report Series*, Vol. 2020, pp. 1–16, URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ets2.12294>, DOI: <http://dx.doi.org/https://doi.org/10.1002/ets2.12294>.
- Quah, Danny (1990) “An improved rate for non-negative definite consistent covariance matrix estimation with heterogeneous dependent data,” *Economics Letters*, Vol. 33, pp. 133–140, URL: <https://ideas.repec.org/a/eee/ecolet/v33y1990i2p133-140.html>.

- Robertson, John C, Ellis W Tallman, and Charles H Whiteman (2005) “Forecasting Using Relative Entropy,” *Journal of Money, Credit and Banking*, Vol. 37, pp. 383–401, URL: <https://ideas.repec.org/a/mcb/jmoncb/v37y2005i3p383-401.html>.
- Romer, Christina D. and David H. Romer (2000) “Federal Reserve Information and the Behavior of Interest Rates,” *American Economic Review*, Vol. 90, pp. 429–457, URL: <https://www.aeaweb.org/articles?id=10.1257/aer.90.3.429>, DOI: <http://dx.doi.org/10.1257/aer.90.3.429>.
- Saiz, Albert (2010) “The geographic determinants of housing supply,” *The Quarterly Journal of Economics*, Vol. 125, pp. 1253–1296.
- Sims, Christopher A. (1980) “Macroeconomics and Reality,” *Econometrica*, Vol. 48, pp. 1–48, URL: <http://www.jstor.org/stable/1912017>.
- Stock, James H and Mark W Watson (2002) “Macroeconomic Forecasting Using Diffusion Indexes,” *Journal of Business & Economic Statistics*, Vol. 20, pp. 147–162, URL: <https://doi.org/10.1198/073500102317351921>, DOI: <http://dx.doi.org/10.1198/073500102317351921>.
- Székely, Gábor J. and Nail K. Bakirov (2003) “Extremal probabilities for Gaussian quadratic forms,” *Probability Theory and Related Fields*, Vol. 126, pp. 184–202.
- Tallman, Ellis W. and Saeed Zaman (2020) “Combining survey long-run forecasts and nowcasts with BVAR forecasts using relative entropy,” *International Journal of Forecasting*, Vol. 36, pp. 373–398, URL: <https://ideas.repec.org/a/eee/intfor/v36y2020i2p373-398.html>, DOI: <http://dx.doi.org/10.1016/j.ijforecast.2019.1016/j.ijforecast.2019>.
- Tallman, Emily and Mike West (2022) “On Entropic Tilting and Predictive Conditioning.”
- Theil, H. (1963) “On the Use of Incomplete Prior Information in Regression Analysis,” *Journal of the American Statistical Association*, Vol. 58, pp. 401–414, URL: <http://www.jstor.org/stable/2283275>.
- Wang, Ye, Cyrus Samii, Haoge Chang, and P. M. Aronow (2023) “Design-Based Inference for Spatial Experiments under Unknown Interference.”
- Winberry, Thomas (2018) “A method for solving and estimating heterogeneous agent macro models,” *Quantitative Economics*, Vol. 9, pp. 1123–1151, URL: <http://qeconomics.org/ojs/index.php/qe/article/view/617>, DOI: <http://dx.doi.org/10.3982/QE740>.

Yoon, Jaehyun (2021) “Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach,” *Computational Economics*, Vol. 57, pp. 247–265, URL: https://ideas.repec.org/a/kap/compec/v57y2021i1d10.1007_s10614-020-10054-w.html, DOI: <http://dx.doi.org/10.1007/s10614-020-10054->.

Zigler, Corwin M. and Georgia Papadogeorgou (2021) “Bipartite Causal Inference with Interference,” *Statistical Science*, Vol. 36, pp. 109 – 123, URL: <https://doi.org/10.1214/19-STS749>, DOI: <http://dx.doi.org/10.1214/19-STS749>.

Appendix A

Appendix for Chapter 1

A.1 Spatial Point Processes and Analytical Example

A.1.1 Poisson Processes

Definition 29. (Poisson Point Process) Suppose there exists an intensity function $\lambda : \mathbb{R}^d \rightarrow [0, \infty)$ where

1. For any bounded $B \subset \mathbb{R}^d$,

$$\int_B \lambda(x) dx < \infty$$

2. For every collection of disjoint sets $B_1, \dots, B_k \subset \mathbb{R}^d$, letting $N(B_i)$ denote the number of elements in region B_i , $N(B_i) \perp N(B_j)$ and

$$P[N(B_i) = n_i] \sim \text{Poisson}(\Lambda(B_i))$$

where

$$\Lambda(B_i) = \int_{B_i} \lambda(x) dx$$

Then this defines a Poisson point process over the spatial domain $\mathbb{R}^d$. See [Diggle \(2013\)](#) for a textbook reference.

A.1.2 Proofs for Section 1.2.5

This section derives the treatment probabilities for the simple example stated in Proposition 1. The proof is done in two pieces: first deriving treatment probabilities when the assignment mechanism is a binomial point process. It is then straightforward to derive treatment probabilities for the Poisson case.

Definition 30. (Binomial Assignment Mechanism) The binomial point process on circular domain S fixes the number of treatment centers C and scatters them independently and uniformly over the domain S .

The difference from the Poisson assignment in Section 1.2.5 is that the Poisson process draws the number of treatment centers C . A well-known result from the spatial statistics literature is that a Poisson process conditioned on the number of treatment centers C is a binomial process (Diggle, 2013).

Throughout, let $B_\delta(\ell)$ denote a ball of size δ about some location $\ell \in S$, while $|A|$ denotes the size of some closed connected subset of S . For S representing a circle of circumference $|S|$, $|A|$ denotes the arc length of the connected subset $A \subset S$.

Lemma 3. Assuming the same setup as in Section 1.2.5 but using a binomial point process instead for the assignment mechanism, treatment probabilities are

$$\begin{aligned}\pi_i &= P[D_i = 1|C] = 1 - (1 - 2\delta/|S|)^C \\ \pi_{ij} &= P[D_i = 1, D_j = 1|C] = \begin{cases} 1 - 2\left(1 - \frac{2\delta}{|S|}\right)^C + \left(1 - \frac{4\delta}{|S|}\right)^C & B_\delta(\ell_i) \cap B_\delta(\ell_j) = \emptyset \\ 1 - 2\left(1 - \frac{2\delta}{|S|}\right)^C + \left(1 - \frac{2\delta + |B_\delta(\ell_j) \setminus B_\delta(\ell_i)|}{|S|}\right)^C & B_\delta(\ell_i) \cap B_\delta(\ell_j) \neq \emptyset \end{cases}\end{aligned}$$

Proof. Under a fixed number C of treatment centers that are independently and uniformly scattered over the circle and treatment rule given by Expression Equation (1.3),

$$\begin{aligned}P[D_i = 1|C] &= P[\text{At least one treatment center in } B_\delta(\ell_i)] \\ &= 1 - P[\text{No treatment centers in } B_\delta(\ell_i)] \\ &= 1 - \left(1 - P[\text{Any one treatment center in } B_\delta(\ell_i)]\right)^C \\ &= 1 - \left(1 - |B_\delta(\ell_i)| / |S|\right)^C \\ &= 1 - (1 - 2\delta / |S|)^C\end{aligned}$$

There are two possibilities for second order probabilities for distinct units i and j :

- $B_\delta(\ell_i) \cap B_\delta(\ell_j) = \emptyset$, no overlap: Then, using conditioning and the assumptions of the binomial assignment mechanism,

$$\begin{aligned}
& P[D_i = 1, D_j = 1 \mid C] \\
&= P[D_j = 1 \mid D_i = 1, C] \cdot P[D_i = 1 \mid C] \\
&= P[D_j = 1] - \left(1 - \left(1 - \frac{|B_\delta(\ell_j)|}{|S| - |B_\delta(\ell_i)|}\right)^C\right) \cdot P[D_i = 0] \\
&= \left[1 - \left(1 - \frac{|B_\delta(\ell_j)|}{|S|}\right)^C\right] - \left(1 - \left(1 - \frac{|B_\delta(\ell_j)|}{|S| - |B_\delta(\ell_i)|}\right)^C\right) \left(1 - \frac{|B_\delta(\ell_i)|}{|S|}\right)^C \\
&= 1 - \left(\frac{|S| - |B_\delta(\ell_j)|}{|S|}\right)^C - \left(\frac{|S| - |B_\delta(\ell_i)|}{|S|}\right)^C + \left(\frac{|S| - |B_\delta(\ell_i)| - |B_\delta(\ell_j)|}{|S|}\right)^C \\
&= 1 - \left(1 - \frac{2\delta}{|S|}\right)^C - \left(1 - \frac{2\delta}{|S|}\right)^C + \left(1 - \frac{4\delta}{|S|}\right)^C
\end{aligned}$$

- $B_\delta(\ell_i) \cap B_\delta(\ell_j) \neq \emptyset$, overlap: Using conditioning and the assumptions of the binomial assignment mechanism,

$$\begin{aligned}
& P[D_i = 1, D_j = 1 \mid C] \\
&= P[D_j = 1] - (1 - (1 - P[\text{Any one in } B_\delta(\ell_j) \mid \text{None in } B_\delta(\ell_i)])^C) \cdot P[D_i = 0] \\
&= P[D_j = 1] - \left(1 - \left(1 - \frac{|B_\delta(\ell_j) \setminus B_\delta(\ell_i)|}{|S| - |B_\delta(\ell_i)|}\right)^C\right) \cdot P[D_i = 0] \\
&= 1 - \left(1 - \frac{|B_\delta(\ell_j)|}{|S|}\right)^C - \left(1 - \left(1 - \frac{|B_\delta(\ell_j) \setminus B_\delta(\ell_i)|}{|S| - |B_\delta(\ell_i)|}\right)^C\right) \left(1 - \frac{|B_\delta(\ell_i)|}{|S|}\right)^C \\
&= 1 - \left(\frac{|S| - |B_\delta(\ell_j)|}{|S|}\right)^C - \left(\frac{|S| - |B_\delta(\ell_i)|}{|S|}\right)^C + \left(\frac{|S| - |B_\delta(\ell_i)| - |B_\delta(\ell_j) \setminus B_\delta(\ell_i)|}{|S|}\right)^C \\
&= 1 - \left(1 - \frac{2\delta}{|S|}\right)^C - \left(1 - \frac{2\delta}{|S|}\right)^C + \left(1 - \frac{2\delta + |B_\delta(\ell_j) \setminus B_\delta(\ell_i)|}{|S|}\right)^C
\end{aligned}$$

□

Having computed treatment probabilities under a binomial point process assignment mechanism that fixes the number of treatment centers C , I can now derive treatment probabilities under a Poisson point process. The difference is that now $C \sim \text{Poisson}(\lambda|S|)$.

Proof. (Proposition 1) Conditional on C , the treatment probabilities are those of the binomial point process

assignment mechanism. Therefore,

$$P[D_i = 1] = \sum_{c=0}^{\infty} P[D_i = 1 \mid C = c] P[C = c]$$

Substituting in the expressions derived above for the binomial point process,

$$P[D_i = 1] = \sum_{c=0}^{\infty} \left[1 - \left(1 - \frac{2\delta}{|S|} \right)^c \right] \cdot \frac{(\lambda|S|)^c}{c!} \exp\{-\lambda|S|\}$$

Break up the term in brackets and rewrite

$$P[D_i = 1] = \sum_{c=0}^{\infty} \frac{(\lambda|S|)^c}{c!} \exp\{-\lambda|S|\} - \sum_{c=0}^{\infty} \left(1 - \frac{2\delta}{|S|} \right)^c \cdot \frac{(\lambda|S|)^c}{c!} \exp\{-\lambda|S|\}$$

Notice that the first sum is a sum over $P[X = c]$ where $X \sim \text{Poisson}(\lambda|S|)$, hence

$$P[D_i = 1] = 1 - \sum_{c=0}^{\infty} \frac{1}{c!} \left[\left(1 - \frac{2\delta}{|S|} \right) \cdot \lambda|S| \right]^c e^{-\lambda|S|}$$

Rewriting again,

$$\begin{aligned} P[D_i = 1] &= 1 - \sum_{c=0}^{\infty} \frac{1}{c!} \left[\left(1 - \frac{2\delta}{|S|} \right) \cdot \lambda|S| \right]^c e^{-\lambda|S|(1-\frac{2\delta}{|S|})} e^{-\lambda|S|[1-(1-\frac{2\delta}{|S|})]} \\ &= 1 - e^{-\lambda \cdot 2\delta} \sum_{c=0}^{\infty} \frac{1}{c!} \cdot \left[\left(1 - \frac{2\delta}{|S|} \right) \cdot \lambda|S| \right]^c e^{-\lambda|S|(1-\frac{2\delta}{|S|})} \end{aligned}$$

But notice that the sum is a sum over $P[X = c]$ where $X \sim \text{Poisson}(\lambda|S|(1 - 2\delta/|S|))$, hence they must all sum to unity, leaving

$$P[D_i = 1] = 1 - e^{-\lambda|B_\delta(\ell_i)|} \cdot 1 = 1 - e^{-\lambda \cdot 2\delta}$$

For the second order inclusion probabilities, the two cases are as follows:

- $B_\delta(\ell_i) \cap B_\delta(\ell_j) = \emptyset$, no overlap: Then

$$\begin{aligned} P[D_i = 1, D_j = 1] &= \sum_{c=0}^{\infty} P[D_i = 1, D_j = 1 \mid C = c] \cdot P[C = c] \\ &= \sum_{c=0}^{\infty} \left[1 - \left(1 - \frac{2\delta}{|S|} \right)^c - \left(1 - \frac{2\delta}{|S|} \right)^c + \left(1 - \frac{4\delta}{|S|} \right)^c \right] \frac{(\lambda|S|)^c e^{-\lambda|S|}}{c!} \end{aligned}$$

Breaking up and simplifying,

$$\begin{aligned}
P[D_i = 1, D_j = 1] &= \sum_{c=0}^{\infty} \frac{(\lambda|S|)^c e^{-\lambda|S|}}{c!} - 2 \sum_{c=0}^{\infty} \frac{(\lambda|S|(1 - 2\delta/|S|))^c e^{-\lambda|S|}}{c!} + \sum_{c=0}^{\infty} \frac{(\lambda|S|(1 - 4\delta/|S|))^c e^{-\lambda|S|}}{c!} \\
&= \sum_{c=0}^{\infty} \frac{(\lambda|S|)^c e^{-\lambda|S|}}{c!} - 2e^{-\lambda|S|(1-2\delta/|S|)} \sum_{c=0}^{\infty} \frac{(\lambda|S|(1 - 2\delta/|S|))^c e^{-\lambda|S|2\delta/|S|}}{c!} \\
&\quad + e^{-\lambda|S|4\delta/|S|} \sum_{c=0}^{\infty} \frac{(\lambda|S|(1 - 4\delta/|S|))^c e^{-\lambda|S|(1-4\delta/|S|)}}{c!} \\
&= 1 - 2e^{-2\delta\lambda} + e^{-4\delta\lambda} = (1 - e^{-2\delta\lambda})^2 = \pi_i^2
\end{aligned}$$

which used that the sums were sums of $P[X = c]$ for Poisson random variables. For these i and j pairs, note that

$$1 - \pi_i - \pi_j + \pi_{ij} = e^{-4\delta\lambda}$$

- $B_\delta(\ell_i) \cap B_\delta(\ell_j) \neq \emptyset$, overlap: Through similar arguments, can show that

$$P[D_i = 1, D_j = 1] = 1 - 2e^{-2\delta\lambda} + e^{-2\delta\lambda - \lambda|B_\delta(\ell_j) \setminus B_\delta(\ell_i)|}$$

For these i and j pairs

$$1 - \pi_i - \pi_j + \pi_{ij} = e^{-2\delta\lambda - \lambda|B_\delta(\ell_j) \setminus B_\delta(\ell_i)|}$$

Substituting in the arc length for $|B_\delta(\ell_j) \setminus B_\delta(\ell_i)|$ gives the expressions of [Section 1.2.5](#). $\square$

A.2 Background for Asymptotic Results

A.2.1 Setup and Metrics

I first set up notation and restate concepts in the notation. This restatement reconciles the different definitions, notation, and assumptions of [Bolthausen \(1982\)](#) and [Jenish and Prucha \(2009\)](#) (the underlying CLTs), as well as [Conley \(1999\)](#) and this paper (which use these CLTs). This setup explicitly allows for triangular arrays as in [Jenish and Prucha \(2009\)](#), which generalizes [Bolthausen \(1982\)](#) and [Conley \(1999\)](#).

Definition 31. (Domain, Units, and Sequence of Populations) $\{\mathcal{I}_n\}$ is a sequence of finite lattices collecting the locations of units within the sequence of spatial domains $\{S_n\}$, i.e. $\mathcal{I}_n \subset S_n$ and $S_n \subseteq \mathbb{R}^2$ for all n . Lattices in the sequence $\{\mathcal{I}_n\}$ are subsets of some countable superdomain lattice $\mathcal{I} \subset \mathbb{R}^2$ with the number of elements $N_n := |\mathcal{I}_n| \rightarrow \infty$ as $n \rightarrow \infty$. All domains may be irregular and all lattices may exhibit uneven spacing. Throughout, let $|U|$ denote the cardinality of a finite subset $U \subseteq \mathcal{I}$.

Definition 32. (Coordinate-wise Distance Metrics) Define the following metric for the distance between any two locations $\ell, \ell' \in \mathbb{R}^2$:

$$\rho(\ell, \ell') = \max_{k \in \{1, 2\}} |\ell_k - \ell'_k|$$

where ℓ_k denotes the k th coordinate of location $\ell \in \mathbb{R}^2$. From a given location ℓ , the set of units that are distance c away are a box of side length $2c$ centered at that location.

And between any two subsets $U, V \subseteq \mathcal{I}$ consisting of sets of unit locations, define

$$\rho(U, V) = \inf\{\rho(\ell, \ell') \mid \ell \in U, \ell' \in V\}$$

This is the smallest distance (coordinate-wise) between any pair of units from the sets.

Note these reflect notation and the distance metric used by [Jenish and Prucha \(2009\)](#). These also reflect the distance metric (but not notation) of [Bolthausen \(1982\)](#), which underlied the proof of [Conley \(1999\)](#).

Definition 33. (*Euclidean Distance Metrics*): Euclidean distance is used in this paper, for the distance between any two locations $\ell, \ell' \in \mathbb{R}^2$:

$$d(\ell, \ell') = \left[\sum_{k=1}^2 (\ell_k - \ell'_k)^2 \right]^{1/2}$$

As usual, from a given location ℓ , the set of units that are distance c away are a sphere of radius c .

Between any two subsets $U, V \subseteq \mathcal{I}$ consisting of sets of unit locations, define

$$d(U, V) = \inf\{d(\ell, \ell') \mid \ell \in U, \ell' \in V\}$$

This is the smallest (Euclidean) distance between any pair of units from the sets.

The next set of results are important for allowing translation between metrics. This is necessary because the the central limit theorems of [Bolthausen \(1982\)](#) and [Jenish and Prucha \(2009\)](#) use the minimum

coordinate-wise distance metrics (denoted ρ), which is more common in the theoretical literature. On the other hand, the assumptions in [Conley \(1999\)](#) and this paper use Euclidean metrics (denoted d).

Lemma 4. *By definition of the metrics*

- (i) $d(\ell, \ell') = c$ implies that $\rho(\ell, \ell') \in [c/\sqrt{2}, c]$
- (ii) $\rho(\ell, \ell') = c$ implies that $d(\ell, \ell') \in [c, \sqrt{2}c]$
- (iii) $\rho(U, V) = c$ implies that $d(U, V) \geq c$.

Proof. (i) and (ii) are straightforward and can be shown geometrically. This leaves

(i) Writing out,

$$\begin{aligned} \rho(U, V) = c &\implies \rho(\ell, \ell') \geq c \quad \forall \ell \in U, \ell' \in V \\ &\implies d(\ell, \ell') \geq c \quad \forall \ell \in U, \ell' \in V \\ &\implies d(U, V) \geq c \end{aligned}$$

where the second line followed from (ii) and the final line from the definition of $d(U, V)$ as the infimum. □

A.2.2 Mixing Definitions and Relationships

I first state the mixing coefficient definition of [Jenish and Prucha \(2009\)](#). It uses the ρ metric and is identical to [Bolthausen \(1982\)](#), except that it accommodates triangular arrays via an outer supremum over n , which indexes the sequence of populations. I then follow with another alternative mixing coefficient definition that uses Euclidean distance instead.

Definition 34. (Mixing Coefficients $\bar{\alpha}_{k,\ell}(m)$) As in [Jenish and Prucha \(2009\)](#) define

$$\bar{\alpha}_{k,\ell}(m) = \sup_n \left[\sup_{|U| \leq k, |V| \leq \ell, \rho(U, V) \geq m} \left[\sup_{A \in \sigma(\{Z_{n,i}\}_{\ell_i \in U}), B \in \sigma(\{Z_{n,i}\}_{\ell_i \in V})} |P[A \cap B] - P[A]P[B]| \right] \right]$$

where $\sigma(A)$ is the σ -algebra generated by A and $U, V \subseteq \mathcal{I}$. This generalizes the time series setting; see [Jenish and Prucha \(2009\)](#) for references. This definition was used in [Bolthausen \(1982\)](#) and [Conley \(1999\)](#) without the outer supremum over n (since those papers did not allow triangular arrays), but is labeled $\alpha_{k,\ell}(m)$ and $\tilde{\alpha}_{k,\ell}(m)$ respectively, in those papers.

Definition 35. (Mixing Coefficients $\alpha_{k,\ell}(m)$) To use Euclidean distance instead, define

$$\alpha_{k,\ell}(m) = \sup_n \left[\sup_{|U| \leq k, |V| \leq \ell, d(U,V) \geq m} \left[\sup_{A \in \sigma(\{Z_{n,i}\}_{\ell_i \in U}), B \in \sigma(\{Z_{n,i}\}_{\ell_i \in V})} |P[A \cap B] - P[A]P[B]| \right] \right]$$

Notice that in the second supremum, the $\rho(U, V)$ was simply replaced with $d(U, V)$. This definition is analogous to $\alpha_{k,\ell}(n)$ in [Conley \(1999\)](#) but adding $\sup_n$ to allow triangular arrays.

Lemma 5. *The mixing coefficients satisfy*

$$\bar{\alpha}_{k,\ell}(m) \leq \alpha_{k,\ell}(m)$$

Proof. From Lemma 4, $\rho(U, V) \geq m$ implies that $d(U, V) \geq m$, implying the set inclusion

$$\{U, V \mid |U| \leq k, |V| \leq \ell, \rho(U, V) \geq m\} \subseteq \{U, V \mid |U| \leq k, |V| \leq \ell, d(U, V) \geq m\}.$$

Therefore, for any given n ,

$$\begin{aligned} & \sup_{|U| \leq k, |V| \leq \ell, \rho(U,V) \geq m} \left[\sup_{A \in \sigma(\{Z_{n,i}\}_{\ell_i \in U}), B \in \sigma(\{Z_{n,i}\}_{\ell_i \in V})} |P[A \cap B] - P[A]P[B]| \right] \\ & \leq \sup_{|U| \leq k, |V| \leq \ell, d(U,V) \geq m} \left[\sup_{A \in \sigma(\{Z_{n,i}\}_{\ell_i \in U}), B \in \sigma(\{Z_{n,i}\}_{\ell_i \in V})} |P[A \cap B] - P[A]P[B]| \right] \end{aligned}$$

Taking supremum over n on the righthand then lefthand side, and then identifying the resulting expressions with the definitions of $\bar{\alpha}_{k,\ell}(m)$ and $\alpha_{k,\ell}(m)$, the result follows. $\square$

A.2.3 Mixing Assumptions for Normality

In order to obtain a CLT and consistent variance estimator, there are several different assumptions that must be made for the mixing coefficients. I state them individually, then state several “bundled” assumptions that correspond to the sets of assumptions used in [Bolthausen \(1982\)](#), [Conley \(1999\)](#), [Jenish and Prucha \(2009\)](#), and this paper. I focus on those assumptions for asymptotic normality only, as the assumptions necessary for consistent variance estimation are the same, plus a strengthening of one assumption, which I stated in the body of the paper.

All assumptions will be stated for triangular arrays as in [Jenish and Prucha \(2009\)](#). For ease of reference, I also state the associated labels for each assumption within [Jenish and Prucha \(2009\)](#) as well as [Conley \(1999\)](#),

with the caveat that Conley did not consider triangular arrays and so there is an implicit generalization rather than perfect match.

Assumption 18. (3(b) in [Jenish and Prucha \(2009\)](#); B1' in [Conley \(1999\)](#)) For $k + \ell \leq 4$,

$$\sum_{m=1}^{\infty} m^{d-1} \cdot \bar{\alpha}_{k,\ell}(m) < \infty$$

where $d = 2$. In words, the dependence between any two units, any unit and pair, any two pairs, or any unit and three units must decay quickly enough with distance.

Assumption 19. (3(c) in [Jenish and Prucha \(2009\)](#); B2' in [Conley \(1999\)](#)) For some $\varepsilon > 0$,

$$\bar{\alpha}_{1,\infty}(m) = O(m^{-2-\varepsilon})$$

In words, the dependence between any given unit and any arbitrarily large set of other units must decay quickly enough with distance.

Assumption 20. (B3' in [Conley \(1999\)](#)) For some $\delta > 0$ and $d = 2$,

$$\sum_{m=1}^{\infty} m^{d-1} \bar{\alpha}_{1,1}(m)^{\delta/(2+\delta)} < \infty$$

Because $\frac{\delta}{2+\delta} < 1$ and $\bar{\alpha}_{1,1}(m) \rightarrow 0$, this is more stringent than [Assumption 18](#) when $k = \ell = 1$.

Assumption 21. For some $\delta > 0$, and $d = 2$,

$$\sum_{m=1}^{\infty} \bar{\alpha}_{1,1}(m) m^{[d(2+\delta)/\delta]-1} = \sum_{m=1}^{\infty} m^{d-1+2d/\delta} \bar{\alpha}_{1,1}(m) < \infty$$

In words, the dependence between any pair of units must decay quickly enough as a function of distance. The second sum expression is provided for easy comparison with [Assumption 18](#). With the extra $2d/\delta$, this condition is indeed more stringent than [Assumption 18](#) when $k = \ell = 1$.

As noted by [Jenish and Prucha \(2009\)](#) (see discussion after Corollary 1), this is a weaker assumption than [Assumption 20](#) used by [Bolthausen \(1982\)](#), [Conley \(1999\)](#), and this paper.

Assumption 22. (Collected Mixing Assumptions for Cor. 1 of [Jenish and Prucha \(2009\)](#))

- (a) [Assumption 18](#)
- (b) [Assumption 19](#)
- (c) [Assumption 21](#)

Assumption 23. (Mixing Assumptions of [Bolthausen \(1982\)](#))

- (a) [Assumption 18](#)
- (b) [Assumption 19](#)
- (c) [Assumption 20](#)

As discussed in [Jenish and Prucha \(2009\)](#), these assumptions are more stringent than those collected in [Assumption 22](#).

Proposition 20. *Assumption 5 implies Assumption 23 implies Assumption 22*

Proof. The proof that [Assumption 5](#) implies [Assumption 23](#) follows immediately from [Lemma 5](#), which gives that $\bar{\alpha}_{k,\ell}(m) \leq \alpha_{k,\ell}(m)$. In other words, it translates the assumptions on $\alpha(\cdot)$ into assumptions on $\bar{\alpha}(\cdot)$ used by [Jenish and Prucha \(2009\)](#).

Next, the argument that [Assumption 23](#) implies [Assumption 22](#) is given in [Jenish and Prucha \(2009\)](#). $\square$

A.2.4 Spatial Central Limit Theorem under Weak Dependence

Theorem 1. (Corollary 1 of [Jenish and Prucha \(2009\)](#)) *Given a triangular sequence of random fields $\{Z_{n,i}\}_{\ell_i \in \mathcal{I}_n}$, suppose*

- (i) *The domain and units satisfy the setup in [Definition 31](#) with $\mathbb{E}[Z_{n,i}] = 0$ for all n, i*
- (ii) *Any pair of elements with locations $\ell, \ell' \in \mathcal{I}$ are located at distances of at least $\rho_0 > 0$ from each other under the metric $\rho(\ell, \ell')$ defined in [Definition 32](#).*
- (iii) *The weak dependence assumptions in [Assumption 22](#) hold.*
- (iv) *There exists an array of positive real constants $\{c_{n,i}\}$ such that, for some $\delta > 0$,*

$$\lim_{k \rightarrow \infty} \sup_n \sup_{i \in \{1, \dots, N_n\}} \mathbb{E}[|Z_{n,i}/c_{n,i}|^{2+\delta} \mathbf{1}\{|Z_{n,i}/c_{n,i}| > k\}] = 0$$

(v) Defining $M_n := \max_{i \in \{1, \dots, N_n\}} c_{n,i}$ and

$$\sigma_n^2 := \text{Var} \left(\sum_{i=1}^{N_n} Z_{n,i} \right)$$

suppose that

$$\liminf_{n \rightarrow \infty} N_n^{-1} M_n^{-2} \sigma_n^2 > 0$$

Then it is the case that

$$\sigma_n^{-1} \sum_{i=1}^{N_n} Z_{n,i} \xrightarrow{d} \mathcal{N}(0, 1)$$

Proof. The conditions and result are an exact reproduction of [Jenish and Prucha \(2009\)](#). □

Remark. For this paper, I focus on Corollary 1 of [Jenish and Prucha \(2009\)](#) rather than their more general Theorem 1. While both need an assumption controlling extreme values of the random variables $Z_{n,i}$, Corollary 1 employs uniform $L_{2+\delta}$ integrability in (iv), whereas their Theorem 1 instead works with the upper-tail quantile function. In the context of causal inference, bounded potential outcomes is a reasonable assumption many are willing to make (and which is maintained in this paper), which is more obviously and directly related to $L_{2+\delta}$ integrability.

A.2.5 Useful Covariance Inequalities using Mixing Coefficients

To show consistency of the variance estimator, I make use of covariance inequalities that bound the covariance by moments and the mixing coefficients. This was used in [Conley \(1999\)](#). I will also use this approach, but must generalize to allow for triangular sequences.

Lemma 6. (Corollary A.2 in [Hall, Heyde, Birnbaum and Lukacs \(2014\)](#)) *For any random variables X and Y that are $\mathcal{G}$ - and $\mathcal{H}$ -measurable such that $\mathbb{E}|X|^p < \infty$ and $\mathbb{E}|Y|^q$ where $p, q > 1$ and $1/p + 1/q < 1$,*

$$|\text{Cov}(X, Y)| \leq 8 \|X\|_p \cdot \|Y\|_q \cdot a(\mathcal{G}, \mathcal{H})^{1 - \frac{1}{p} - \frac{1}{q}}$$

$$\text{where } a(\mathcal{G}, \mathcal{H}) = \sup_{G \in \mathcal{G}, H \in \mathcal{H}} |P[G \cap H] - P[G]P[H]|$$

where I use a to avoid confusion with α and related mixing coefficients defined previously.

Lemma 7. For any random field $\{Z_{n,\ell_i}\}_{\ell_i \in \mathcal{I}_n}$ in the triangular sequence of random fields (i.e. for any fixed n), if $2 + \delta$ moments exist, then

$$|\text{Cov}(Z_{n,\ell_i}, Z_{n,\ell_j})| \leq 8 \cdot \alpha_{1,1,n}(d(\ell_i, \ell_j))^{\frac{\delta}{2+\delta}} \sup_{\ell_i \in \mathcal{I}_n} \|Z_{n,\ell_i}\|_{2+\delta}^2$$

where

$$\alpha_{k,p,n}(m) := \sup_{|U| \leq k, |V| \leq p, d(U,V) \geq m} \left[\sup_{A \in \sigma(\{Z_{n,i}\}_{\ell_i \in U}), B \in \sigma(\{Z_{n,i}\}_{\ell_i \in V})} |P[A \cap B] - P[A]P[B]| \right]$$

Note that $\alpha_{k,\ell,n}(m)$ is just $\alpha_{k,\ell}(m)$ from [Definition 34](#), but without the leading $\sup_n$.

Remark. This is essentially Lemma 1 of [Bolthausen \(1982\)](#), but because that paper assumed stationarity and non-triangular structure, I will go back to Lemma 6 in order to justify this lemma on grounds closer to first principles.

Proof. Let $m > 0$ and n be given and fixed. Let Z_{n,ℓ_i} and Z_{n,ℓ_j} be any two elements of the random field $\{Z_{n,\ell_i}\}_{\ell_i \in \mathcal{I}_n}$. Apply Lemma 6 to these two random variables, taking $p = q = 2 + \delta$:

$$|\text{Cov}(Z_{n,\ell_i}, Z_{n,\ell_j})| \leq 8 \|Z_{n,\ell_i}\|_{2+\delta} \cdot \|Z_{n,\ell_j}\|_{2+\delta} \cdot a(\sigma(Z_{n,\ell_i}), \sigma(Z_{n,\ell_j}))^{\frac{\delta}{2+\delta}} \quad (\text{A.1})$$

$$\text{where } a(\sigma(Z_{n,\ell_i}), \sigma(Z_{n,\ell_j})) = \sup_{A \in \sigma(Z_{n,\ell_i}), B \in \sigma(Z_{n,\ell_j})} |P[A \cap B] - P[A]P[B]|$$

From here, note that

$$\begin{aligned} & a(\sigma(Z_{n,\ell_i}), \sigma(Z_{n,\ell_j})) \\ &= \sup_{A \in \sigma(Z_{n,\ell_i}), B \in \sigma(Z_{n,\ell_j})} |P[A \cap B] - P[A]P[B]| \\ &\leq \underbrace{\sup_{|U| \leq 1, |V| \leq 1, d(U,V) \geq d(\ell_i, \ell_j)} \left[\sup_{A \in \sigma(\{Z_{n,i}\}_{\ell_i \in U}), B \in \sigma(\{Z_{n,i}\}_{\ell_i \in V})} |P[A \cap B] - P[A]P[B]| \right]}_{\alpha_{1,1,n}(d(\ell_i, \ell_j))} \end{aligned}$$

Using this inequality and definition in Expression [A.1](#), the Lemma follows. $\square$

Lemma 8. (Generalization of Inequality A in [Conley \(1999\)](#)) Suppose that for any fixed n and any $\ell \in \mathcal{I}_n$, $2 + \delta$ moment exist, i.e.

$$\|Z_{n,\ell}\|_{2+\delta} < \infty \quad \forall \ell \in \mathcal{I}_n$$

Then for any pair of random variables Z_{n,ℓ_i} and Z_{n,ℓ_j} in the triangular sequence,

$$\begin{aligned} |\text{Cov}(Z_{n,\ell_i}, Z_{n,\ell_j})| &\leq 8 \cdot \alpha_{1,1}(d(\ell_i, \ell_j))^{\frac{\delta}{2+\delta}} \sup_{\ell \in \mathcal{I}_n} \|Z_{n,\ell}\|_{2+\delta}^2 \\ &\leq 8 \cdot \alpha_{\infty,\infty}(d(\ell_i, \ell_j))^{\frac{\delta}{2+\delta}} \sup_{\ell \in \mathcal{I}_n} \|Z_{n,\ell}\|_{2+\delta}^2 \end{aligned}$$

Proof. Because $2 + \delta$ moments exists, the assumptions of Lemma 7 are satisfied for any fixed n . Hence, for a given n ,

$$\begin{aligned} |\text{Cov}(Z_{n,\ell_i}, Z_{n,\ell_j})| &\leq 8 \cdot \alpha_{1,1,n}(d(\ell_i, \ell_j))^{\frac{\delta}{2+\delta}} \sup_{\ell \in \mathcal{I}_n} \|Z_{n,\ell}\|_{2+\delta}^2 \\ &\leq 8 \cdot \left[\sup_n \alpha_{1,1,n}(d(\ell_i, \ell_j)) \right]^{\frac{\delta}{2+\delta}} \sup_{\ell \in \mathcal{I}_n} \|Z_{n,\ell}\|_{2+\delta}^2 \end{aligned}$$

Then, by Definition 34 for $\alpha_{k,p}(m)$ and the definition of $\alpha_{k,p,n}(m)$ in Lemma 7, note that $\alpha_{1,1}(m) = \sup_n \alpha_{1,1,n}(m)$, implying

$$|\text{Cov}(Z_{n,\ell_i}, Z_{n,\ell_j})| \leq 8 \cdot \alpha_{1,1}(d(\ell_i, \ell_j))^{\frac{\delta}{2+\delta}} \sup_{\ell \in \mathcal{I}_n} \|Z_{n,\ell}\|_{2+\delta}^2 \quad (\text{A.2})$$

From here, what's left is to note that $\alpha_{1,1}(m) \leq \alpha_{\infty,\infty}(m)$. To prove, note for any fixed n ,

$$\begin{aligned} &\sup_{|U| \leq 1, |V| \leq 1, d(U,V) \geq m} \left[\sup_{A \in \sigma(\{Z_{n,\ell_i}\}_{\ell_i \in U}), B \in \sigma(\{Z_{n,\ell_i}\}_{\ell_i \in V})} |P[A \cap B] - P[A]P[B]| \right] \\ &\leq \sup_{|U| \leq \infty, |V| \leq \infty, d(U,V) \geq m} \left[\sup_{A \in \sigma(\{Z_{n,\ell_i}\}_{\ell_i \in U}), B \in \sigma(\{Z_{n,\ell_i}\}_{\ell_i \in V})} |P[A \cap B] - P[A]P[B]| \right] \end{aligned}$$

because the latter is a larger space to search over. Taking supremum over n on the righthand side, then over the left, and identifying the resulting expressions as $\alpha_{1,1}(m)$ and $\alpha_{\infty,\infty}(m)$, the inequality holds, which can be used in Expression A.2 to derive the final result. $\square$

A.3 Proofs for Section 1.3

A.3.1 Useful Lemmas for Normality and Variance Consistency Proofs

Lemma 9. (Crude Bound on Units in a Ball under Expanding Domain Asymptotics) *Let $B(\ell, \delta)$ denote a ball centered at $\ell \in \mathbb{R}^2$ with some finite Euclidean distance/radius δ . Let $\{\ell_i\}_{i \in \mathcal{I}} \subseteq \mathbb{R}^2$ denote an arbitrary countable set of locations such that $d(\ell_i, \ell_k) > d_0$ for any pair $\ell_i \neq \ell_k$ in $\mathcal{I}$, which is the basic assumption for expanding domain asymptotics. Then*

$$|\{\ell_i \mid \ell_i \in B(\ell, \delta)\}| \leq \lceil 2\delta/d_0 \rceil^2 \leq 8(\delta/d_0)^2 \quad \forall \ell \in \mathbb{R}^2, \delta > 0$$

Proof. A square of side length 2δ centered at ℓ will cover $B(\ell, \delta)$. The largest number of units that can fit in the square corresponds to gridded spacing, and so $\lceil 2\delta/d_0 \rceil$ units in a given direction of the grid. This implies $\lceil 2\delta/d_0 \rceil^2$ units in the square containing $B(\ell, \delta)$, which exceeds the amount of units that fit into the “covered” circle.

The final inequality follows by simply doubling the ceiling-based count. This is a very crude inequality, but it is obviously satisfied, dispenses with the ceiling operator, and—most importantly—leaves the ratio $2\delta/d_0$ in place, which will mean certain rates are not affected in the convergence proofs since we are just rescaling by a fixed constant. $\square$

Lemma 10. (Bound on Double Sum of Covariances) *Assumption 5 plus Assumption 4 imply*

$$\frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |\text{Cov}(D_{n,i}, D_{n,j})| = O(1)$$

Proof. By Part (iii) of Assumption 5, it must be the case that $\alpha_{1,1}(m)^{\delta/(2+\delta)} = O(m^{-2-c})$ for some constant $c > 0$, since that ensures

$$\sum_{m=1}^{\infty} m^1 O(m^{-2-c}) = \sum_{m=1}^{\infty} O(1/m^{1+c}) < \infty$$

Therefore, together with Lemma 8 and the fact that $D_{n,i} \leq 1$,

$$\frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |\text{Cov}(D_{n,i}, D_{n,j})| \leq \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} 8\alpha_{1,1}(d(\ell_i, \ell_j))^{\delta/(2+\delta)} \leq \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} 8k \frac{1}{d(\ell_i, \ell_j)^{2+c}}$$

To show this last expression is $O(1)$, I argue that one of the sums converges, hence the double sum will be of the form $N_n \cdot X$ where X is a constant, which divided by N_n yields $O(1)$.

To argue this, note that under the constraint that all units must be separated by distance d_{min} , because of the convexity of $1/x^{2+c}$, the largest possible value for the double sum comes under the densest packing of units. It has been shown that the densest packing is a hexagonal arrangement. Under this hexagonal arrangement, given a generic unit i , all other units fall in concentric circles around the unit. At most, there are 6 units in the first circle, 12 in the second, and $6r$ in the r th concentric ring. Moreover, the distance from unit i to units in the r th ring is rd_{min} . Therefore I can rewrite for any generic unit i ,

$$\sum_{j=1}^{N_n} 8k \frac{1}{d(\ell_i, \ell_j)^{2+c}} \leq \sum_{r=1}^{\infty} 8k \cdot 6r \frac{1}{(rd_{min})^{2+c}} = \sum_{r=1}^{\infty} 8k \cdot 6 \frac{1}{(rd_{min})^{1+c}} = X < \infty$$

□

A.3.2 Proofs for Section 1.3.1

Proof of Lemma 1: WLS Equivalent to HT Given

$$X_{n,i} = \begin{pmatrix} 1 - D_{n,i} \\ D_{n,i} \end{pmatrix} \quad \omega_{n,i} = \frac{D_{n,i}}{\pi_{n,i}} + \frac{1 - D_{n,i}}{1 - \pi_{n,i}}$$

the weighted least squares estimator is

$$\begin{aligned} \hat{\beta}_{n,WLS} &= (X_n' \omega_n X_n)^{-1} X_n' \omega_n Y_n \\ &= (X_n' \omega_n X_n / N_n)^{-1} \frac{1}{N_n} X_n' \omega_n Y_n \\ &= \left[\frac{1}{N_n} \sum_{i=1}^{N_n} \omega_{n,i} \begin{pmatrix} 1 - D_{n,i} & 0 \\ 0 & D_{n,i} \end{pmatrix} \right]^{-1} \left[\frac{1}{N_n} \sum_{i=1}^{N_n} \omega_{n,i} \begin{pmatrix} (1 - D_{n,i}) Y_{n,i} \\ D_{n,i} Y_{n,i} \end{pmatrix} \right] \\ &= \begin{pmatrix} \frac{1}{N_n} \sum_{i=1}^{N_n} \frac{1 - D_{n,i}}{1 - \pi_{n,i}} & 0 \\ 0 & \frac{1}{N_n} \sum_{i=1}^{N_n} \frac{D_{n,i}}{\pi_{n,i}} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{N_n} \sum_{i=1}^{N_n} \frac{1 - D_{n,i}}{1 - \pi_{n,i}} Y_{n,i} \\ \frac{1}{N_n} \sum_{i=1}^N \frac{D_{n,i}}{\pi_{n,i}} Y_{n,i} \end{pmatrix} \\ &= \begin{pmatrix} \hat{R}_{n,0} & 0 \\ 0 & \hat{R}_{n,1} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\mu}_{n,Y(0)} \\ \hat{\mu}_{n,Y(1)} \end{pmatrix} \end{aligned}$$

If $\widehat{R}_{n,1} \neq 0$ and $\widehat{R}_{n,1} \neq 0$, then $\widehat{\beta}_{n, WLS}$ is well-defined and premultiplying by the vector $\begin{pmatrix} -1 & 1 \end{pmatrix}$ gives precisely $\widehat{\alpha}_{n, ATE}$.

A.3.3 Proofs for Section 1.3.2

Proof of Proposition 2: Consistency of Horvitz-Thompson Building Blocks The proof strategy will be to show that the Horvitz Thompson building blocks of $\widehat{\alpha}_{n, ATE}$ converge in mean square—hence in probability, as well—to some target.

Before starting, given that $Y_{n,i} = Y_{n,i}(1)D_{n,i} + Y_{n,i}(0)(1 - D_{n,i})$, define

$$\mu_{n,Y(d)} = \mathbb{E}_n[\widehat{\mu}_{n,Y(d)}] = \frac{1}{N_n} \sum_{i=1}^{N_n} Y_{n,i}(d).$$

I will show that

$$\begin{bmatrix} \widehat{\mu}_{n,Y(1)} - \mu_{n,Y(1)} \\ \widehat{\mu}_{n,Y(0)} - \mu_{n,Y(0)} \\ \widehat{R}_{n,1} \\ \widehat{R}_{n,0} \end{bmatrix} \xrightarrow{m.s.} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

First, by definition of the estimators, identity $Y_{n,i} = Y_{n,i}(1)D_{n,i} + Y_{n,i}(0)(1 - D_{n,i})$, and linearity of the expectation operator

$$\begin{bmatrix} \widehat{\mu}_{n,Y(1)} - \mu_{n,Y(1)} \\ \widehat{\mu}_{n,Y(0)} - \mu_{n,Y(0)} \\ \widehat{R}_{n,1} \\ \widehat{R}_{n,0} \end{bmatrix} = \frac{1}{N_n} \sum_{i=1}^{N_n} \begin{pmatrix} \frac{D_{n,i}}{\pi_{n,i}} Y_{n,i} - Y_{n,i}(1) \\ \frac{1-D_{n,i}}{1-\pi_{n,i}} Y_{n,i} - Y_{n,i}(0) \\ \frac{D_{n,i}}{\pi_{n,i}} \\ \frac{1-D_{n,i}}{1-\pi_{n,i}} \end{pmatrix} = \frac{1}{N_n} \sum_{i=1}^{N_n} \begin{pmatrix} \frac{D_{n,i} - \pi_{n,i}}{\pi_{n,i}} Y_{n,i}(1) \\ \frac{\pi_{n,i} - D_{n,i}}{1-\pi_{n,i}} Y_{n,i}(0) \\ \frac{D_{n,i}}{\pi_{n,i}} \\ \frac{1-D_{n,i}}{1-\pi_{n,i}} \end{pmatrix}.$$

This last form allow easy computation of the first moment, for any n :

$$\mathbb{E}_n \begin{bmatrix} \widehat{\mu}_{n,Y(1)} - \mu_{n,Y(1)} \\ \widehat{\mu}_{n,Y(0)} - \mu_{n,Y(0)} \\ \widehat{R}_{n,1} \\ \widehat{R}_{n,0} \end{bmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}.$$

Convergence in mean square holds if the variance goes to zero,

$$\begin{aligned} & \text{Var}_n \left(\begin{pmatrix} \hat{\mu}_{n,Y(1)} - \mu_{n,Y(1)} & \hat{\mu}_{n,Y(0)} - \mu_{n,Y(0)} & \hat{R}_{n,1} & \hat{R}_{n,0} \end{pmatrix}' \right) \\ &= \frac{1}{N_n^2} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \begin{pmatrix} \frac{Y_{n,i}(1)}{\pi_{n,i}} \frac{Y_{n,j}(1)}{\pi_{n,j}} & -\frac{Y_{n,i}(1)}{\pi_{n,i}} \frac{Y_{n,j}(0)}{1-\pi_{n,j}} & \frac{Y_{n,i}(1)}{\pi_{n,i}} \frac{1}{\pi_{n,j}} & -\frac{Y_{n,i}(1)}{\pi_{n,i}} \frac{1}{1-\pi_{n,j}} \\ -\frac{Y_{n,i}(0)}{1-\pi_{n,i}} \frac{Y_{n,j}(1)}{\pi_{n,j}} & \frac{Y_{n,i}(0)}{1-\pi_{n,i}} \frac{Y_{n,j}(0)}{1-\pi_{n,j}} & -\frac{Y_{n,i}(0)}{1-\pi_{n,i}} \frac{1}{\pi_{n,j}} & \frac{Y_{n,i}(0)}{1-\pi_{n,i}} \frac{1}{1-\pi_{n,j}} \\ \frac{1}{\pi_{n,i}} \frac{Y_{n,j}(1)}{\pi_{n,j}} & -\frac{1}{\pi_{n,i}} \frac{Y_{n,j}(0)}{1-\pi_{n,j}} & \frac{1}{\pi_{n,i}} \frac{1}{\pi_{n,j}} & -\frac{1}{\pi_{n,i}} \frac{1}{1-\pi_{n,j}} \\ -\frac{1}{1-\pi_{n,i}} \frac{Y_{n,j}(1)}{\pi_{n,j}} & \frac{1}{1-\pi_{n,i}} \frac{Y_{n,j}(0)}{1-\pi_{n,j}} & -\frac{1}{1-\pi_{n,i}} \frac{1}{\pi_{n,j}} & \frac{1}{1-\pi_{n,i}} \frac{1}{1-\pi_{n,j}} \end{pmatrix} \text{Cov}(D_{n,i}, D_{n,j}). \end{aligned}$$

The $k\ell$ element of the above variance-covariance matrix can be written generically as follows

$$[\text{Var}_n]_{k\ell} = \frac{1}{N_n^2} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} a_{n,ij,k\ell} \text{Cov}_n(D_{n,i}, D_{n,j}) \leq \frac{1}{N_n^2} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |a_{n,ij,k\ell}| \cdot |\pi_{n,ij} - \pi_{n,i}\pi_{n,j}|$$

where $a_{n,ij,k\ell}$ denotes the $k\ell$ entry of the matrix in the ij th term of the double sum.

Given Assumptions 1 and 2, can bound the coefficients for all n and i, j, k, ℓ :

$$|a_{n,ij,k\ell}| \leq \max\{M_*/\varepsilon, M_*/(1-\varepsilon)\}^2 =: K^2$$

where $M_* = \max\{M, 1\}$. Therefore, for any element $[\text{Var}_n]_{k\ell}$,

$$[\text{Var}_n]_{k\ell} \leq K^2 \frac{1}{N_n^2} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |\pi_{n,ij} - \pi_{n,i}\pi_{n,j}| = o(1)$$

with this asymptotic behavior following from Assumption 3.

The above results together imply joint convergence in probability (via mean square):

$$\begin{aligned} \forall n \quad \mathbb{E}_n \begin{pmatrix} \hat{\mu}_{n,Y(1)} - \mu_{n,Y(1)} \\ \hat{\mu}_{n,Y(0)} - \mu_{n,Y(0)} \\ \hat{R}_{n,1} \\ \hat{R}_{n,0} \end{pmatrix} &= \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \\ \lim_{n \rightarrow \infty} \text{Var}_n \begin{pmatrix} \hat{\mu}_{n,Y(1)} - \mu_{n,Y(1)} \\ \hat{\mu}_{n,Y(0)} - \mu_{n,Y(0)} \\ \hat{R}_{n,1} \\ \hat{R}_{n,0} \end{pmatrix} &= 0 \end{aligned} \quad \Rightarrow \quad \begin{pmatrix} \hat{\mu}_{n,Y(1)} - \mu_{n,Y(1)} \\ \hat{\mu}_{n,Y(0)} - \mu_{n,Y(0)} \\ \hat{R}_{n,1} \\ \hat{R}_{n,0} \end{pmatrix} \xrightarrow{m.s.} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \Rightarrow \xrightarrow{p} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

Proof of Proposition 2: Consistency of WLS/HT ATE Estimator To make use of convergence in the Horvitz-Thompson building blocks from (ii), rewrite

$$\begin{aligned}
& \hat{\alpha}_{n,ATE} - \alpha_{n,ATE} \\
&= \frac{\hat{\mu}_{n,Y(1)}}{\hat{R}_{n,1}} - \frac{\hat{\mu}_{n,Y(0)}}{\hat{R}_{n,0}} - [\mu_{n,Y(1)} - \mu_{n,Y(0)}] \\
&= \frac{\hat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}}{\hat{R}_{n,1}} + \frac{\mu_{n,Y(1)}}{\hat{R}_{n,1}} - \frac{\hat{\mu}_{n,Y(0)} - \mu_{n,Y(0)}}{\hat{R}_{n,0}} - \frac{\mu_{n,Y(0)}}{\hat{R}_{n,0}} - [\mu_{n,Y(1)} - \mu_{n,Y(0)}] \\
&= \frac{1}{\hat{R}_{n,1}} (\hat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}) - \frac{1}{\hat{R}_{n,0}} (\hat{\mu}_{n,Y(0)} - \mu_{n,Y(0)}) + \mu_{n,Y(1)} \left(\frac{1}{\hat{R}_{n,1}} - 1 \right) - \mu_{n,Y(0)} \left(\frac{1}{\hat{R}_{n,0}} - 1 \right).
\end{aligned}$$

By **Assumption 1**, $\mu_{n,Y(d)} = O(M)$. Next, by the continuous mapping theorem and convergence in probability of the Horvitz-Thompson building blocks, conclude that

$$\hat{\alpha}_{n,ATE} - \alpha_{n,ATE} \xrightarrow{p} 0.$$

A.3.4 Proofs for Section 1.3.3

Proof of Proposition 3(i) The independence assumption of this proposition and the definition of $\alpha_{k,\ell}(m)$ implies that for any $k, \ell \in \{1, 2, 3\}$ and $m > m_*$, $\alpha_{k,\ell}(m) = 0$.

(a) To prove **Assumption 20** but with $\alpha_{k,\ell}$ in place of $\bar{\alpha}_{k,\ell}$, note

$$\sum_{m=1}^{\infty} m^{d-1} \alpha_{1,1}(m)^{\delta/(2+\delta)} = \sum_{m=1}^{m_*} m^{d-1} \alpha_{1,1}(m)^{\delta/(2+\delta)} \leq \sum_{m=1}^{m_*} m^{d-1} < \infty$$

which also used that the mixing coefficients are all bounded by 1.

(b) To prove **Assumption 18** but with $\alpha_{k,\ell}$ in place of $\bar{\alpha}_{k,\ell}$, note

$$\sum_{m=1}^{\infty} m \cdot \alpha_{k,\ell}(m) = \sum_{m=1}^{m_*} m \cdot \alpha_{k,\ell}(m) \leq \sum_{m=1}^{m_*} m < \infty$$

again using the fact that mixing coefficients are all bounded by 1.

(c) If $m > m_*$, then $\alpha_{1,\infty} = 0$. Therefore, for any finite $\varepsilon > 0$, $\alpha_{1,\infty} = O(m^{-2-\varepsilon})$.

Proof of Proposition 3(ii) This follows from Lemma 10 and dividing through by N_n .

Proof of Proposition 4 Asymptotic normality of $\widehat{\alpha}_{n,ATE}$ will follow from applying the CLT of [Jenish and Prucha \(2009\)](#)—their Corollary 1 restated in Theorem 1. To do so, I first set up the convergent sum of random variables. I then verify sufficient conditions for asymptotic normality, before finally connecting this convergent sum to rescaled $\widehat{\alpha}_{n,ATE} - \alpha_{n,ATE}$.

(Write Sum for Applying [Jenish and Prucha \(2009\)](#)) Start by defining

$$\begin{aligned} \sum_{i=1}^{N_n} Z_{n,i} &:= \sum_{i=1}^{N_n} \left[\frac{e_{n,i}(0)}{1 - \pi_{n,i}} + \frac{e_{n,i}(1)}{\pi_{n,i}} \right] [D_{n,i} - \pi_{n,i}] \\ \sigma_n^2 &:= \text{Var} \left(\sum_{i=1}^{N_n} Z_{n,i} \right) = \text{Var} \left(\sum_{i=1}^{N_n} \left[\frac{e_{n,i}(0)}{1 - \pi_{n,i}} + \frac{e_{n,i}(1)}{\pi_{n,i}} \right] [D_{n,i} - \pi_{n,i}] \right) \\ &= \underbrace{\sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{e_{n,i}(0)}{1 - \pi_{n,i}} + \frac{e_{n,i}(1)}{\pi_{n,i}} \right] \left[\frac{e_{n,j}(0)}{1 - \pi_{n,j}} + \frac{e_{n,j}(1)}{\pi_{n,j}} \right] [\pi_{n,ij} - \pi_{n,i}\pi_{n,j}]}_{N_n \cdot V_n} \end{aligned}$$

where V_n is defined in [Definition 7](#). Because this is a sum of *correlated* random treatment indicators, standard CLTs for sums of independent random variables do not apply. Instead, I appeal to the CLT of [Jenish and Prucha \(2009\)](#) for spatially correlated scalar random variables under weak dependence. Therefore, the task lies in verifying that the conditions of [Jenish and Prucha \(2009\)](#) (restated in [Theorem 1](#)) hold under the assumptions employed in this paper.

(Verifying Conditions of [Jenish and Prucha \(2009\)](#))

- (i) The triangular setup matches. Notice also the $Z_{n,i}$ are all mean zero (as required) because the $e_{n,i}(d)$ and treatment probabilities $\pi_{n,i}$ are non-stochastic while $\mathbb{E}[D_{n,i}] = \pi_{n,i}$.
- (ii) By [Assumption 4](#), there is a fixed constant d_{min} independent of n such that $d(\ell_{n,i}, \ell_{n,j}) > d_{min}$ for all n and pairs i, j , where $d(\cdot, \cdot)$ gives distance under the Euclidean metric. Then, by [Lemma 4](#),

$$\rho(\ell_{n,i}, \ell_{n,j}) > d_{min}/\sqrt{2} =: \rho_0, \quad \forall n, i, j \tag{A.3}$$

Hence the condition is satisfied for this value of ρ_0 .

- (iii) The weak dependence conditions in [Assumption 5](#) placed on treatment $D_{n,i}$ pass to the $Z_{n,i}$ defined above, since $Z_{n,i}$ is a simple rescaling and recentering of $D_{n,i}$ by non-stochastic terms. We then recognize that the conditions in [Assumption 5](#) are identical to [Assumption 5](#). From there, apply

Proposition 20 to conclude that the required weak dependence conditions hold.

- (iv) By bounded potential outcomes from Assumption 1, nondegenerate treatment probabilities from Assumption 2, and $D_{n,i} \in \{0, 1\}$,

$$\implies |Z_{n,i}| \leq \left| \frac{e_{n,i}(0)}{1 - \pi_{n,i}} + \frac{e_{n,i}(1)}{\pi_{n,i}} \right| \cdot 1 \leq \frac{|e_{n,i}(0)| + |e_{n,i}(1)|}{\min\{\varepsilon, 1 - \varepsilon\}} \leq \underbrace{\frac{2M}{\min\{\varepsilon, 1 - \varepsilon\}}}_{=:K}$$

where M is from Assumption 1 and ε is from Assumption 2. Hence, $|Z_{n,i}|$ is bounded.

Therefore, to verify (iv) of Theorem 1, take $c_{n,i} = 1$ and note for any k

$$\mathbb{E}[|Z_{n,i}|^{2+\delta} \mathbf{1}\{|Z_{n,i}| > k\}] \leq \mathbb{E}[K^{2+\delta} \mathbf{1}\{|Z_{n,i}| > k\}] = K^{2+\delta} P[|Z_{n,i}| > k]$$

By the bound from above, $P[|Z_{n,i}| > k] = 0$ for any $k > K$. Taking some finite $K^* > K$,

$$\lim_{k \rightarrow \infty} \sup_n \sup_i \mathbb{E}[|Z_{n,i}|^{2+\delta} \mathbf{1}\{|Z_{n,i}| > k\}] \leq \sup_n \sup_i |K|^{2+\delta} \cdot P[|Z_{n,i}| > K^*] = 0$$

- (v) The $Z_{n,i}$ were defined above, and inspecting its definition together with Definition 7, it's clear that $\sigma_n^2 = \text{Var}(\sum_{i=1}^{N_n} Z_{n,i}) = N_n V_n$. From there, because $c_{n,i} = 1$ from (iv), it is the case that $M_n = 1$ and

$$N_n^{-1} \sigma_n^2 = V_n > 0$$

where the last inequality followed from Assumption 6.

Having verified all conditions in Theorem 1, invoke the result to conclude

$$\sigma_n^{-1} \sum_{i=1}^{N_n} Z_{n,i} \xrightarrow{d} \mathcal{N}(0, 1) \iff V_n^{-1/2} \frac{1}{\sqrt{N_n}} \sum_{i=1}^{N_n} \left[\frac{e_{n,i}(0)}{1 - \pi_{n,i}} + \frac{e_{n,i}(1)}{\pi_{n,i}} \right] [D_{n,i} - \pi_{n,i}] \xrightarrow{d} \mathcal{N}(0, 1)$$

(Relating to $\hat{\alpha}_{ATE}$) Having established asymptotic normality for the above quantity, I now connect this to the rescaled and recentered effect estimator. To start, I first make use of the equivalence established in Proposition 1 of $\hat{\alpha}_{n,ATE} = (-1, 1) \hat{\beta}_{n,WLS}$. This equivalence implies that I can write the recentered effect

estimator (in the n th population) as follows

$$\begin{aligned}
\hat{\alpha}_{n,ATE} - \alpha_{n,ATE} &= \begin{pmatrix} -1 \\ 1 \end{pmatrix}' \left[\hat{\beta}_{n,ATE} - \begin{pmatrix} \mu_{n,Y(0)} \\ \mu_{n,Y(1)} \end{pmatrix} \right] \\
&= \begin{pmatrix} -1 \\ 1 \end{pmatrix}' \left[(X_n' \omega_n X_n / N_n)^{-1} \frac{1}{N_n} X_n' \omega_n Y_n - \begin{pmatrix} \mu_{n,Y(0)} \\ \mu_{n,Y(1)} \end{pmatrix} \right] \\
&= \begin{pmatrix} -1 \\ 1 \end{pmatrix}' (X_n' \omega_n X_n / N_n)^{-1} \left[\frac{1}{N_n} X_n' \omega_n Y_n - (X_n' \omega_n X_n / N_n) \begin{pmatrix} \mu_{n,Y(0)} \\ \mu_{n,Y(1)} \end{pmatrix} \right] \\
&= \begin{pmatrix} -1 \\ 1 \end{pmatrix}' \underbrace{\begin{pmatrix} \hat{R}_{n,0} & 0 \\ 0 & \hat{R}_{n,1} \end{pmatrix}^{-1}}_{(a)} \underbrace{\frac{1}{N_n} X_n' \omega_n}_{(b)} \underbrace{\left[Y_n - X_n \begin{pmatrix} \mu_{n,Y(0)} \\ \mu_{n,Y(1)} \end{pmatrix} \right]}_{(c)}
\end{aligned}$$

Note that (a) can be written out of the expression. By the convergence in probability established in Proposition 2(ii) and the continuous mapping theorem,

$$\begin{pmatrix} \hat{R}_{n,0} & 0 \\ 0 & \hat{R}_{n,1} \end{pmatrix}^{-1} = I_2^{-1} + \begin{pmatrix} \hat{R}_{n,0} & 0 \\ 0 & \hat{R}_{n,1} \end{pmatrix}^{-1} - I_2^{-1} = I_2(1 + o_p(1))$$

Therefore,

$$\hat{\alpha}_{n,ATE} - \alpha_{n,ATE} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}' \underbrace{\frac{1}{N_n} X_n' \omega_n}_{(b)} \underbrace{\left[Y_n - X_n \begin{pmatrix} \mu_{n,Y(0)} \\ \mu_{n,Y(1)} \end{pmatrix} \right]}_{(c)} (1 + o_p(1))$$

Now, I develop (b) times (c) into a sum that can be related to the sum that I have already shown to be asymptotically normal. To start, write the i th element of vector (c):

$$\begin{aligned}
Y_{n,i} - X_{n,i}' \begin{pmatrix} \mu_{n,Y(0)} \\ \mu_{n,Y(1)} \end{pmatrix} &= Y_{n,i}(1)D_{n,i} + Y_{n,i}(0)(1 - D_{n,i}) - [(1 - D_{n,i})\mu_{n,Y(0)} + D_{n,i}\mu_{n,Y(1)}] \\
&= D_{n,i}[Y_{n,i}(1) - \mu_{n,Y(1)}] + (1 - D_{n,i})[Y_{n,i}(0) - \mu_{n,Y(0)}]
\end{aligned}$$

Use the above to rewrite the quantity (b) times (c) as follows,

$$\begin{aligned}
& \frac{1}{N_n} X_n' \omega_n \left[Y_n - X_n \begin{pmatrix} \mu_{n,Y(0)} \\ \mu_{n,Y(1)} \end{pmatrix} \right] \\
&= \frac{1}{N_n} \sum_{i=1}^{N_n} \begin{pmatrix} 1 - D_{n,i} \\ D_{n,i} \end{pmatrix} \left[\frac{D_{n,i}}{\pi_{n,i}} + \frac{1 - D_{n,i}}{1 - \pi_{n,i}} \right] [D_{n,i}[Y_{n,i}(1) - \mu_{n,Y(1)}] + (1 - D_{n,i})[Y_{n,i}(0) - \mu_{n,Y(0)}]] \\
&= \frac{1}{N_n} \sum_{i=1}^{N_n} \begin{pmatrix} \frac{1 - D_{n,i}}{1 - \pi_{n,i}} [Y_{n,i}(0) - \mu_{n,Y(0)}] \\ \frac{D_{n,i}}{\pi_{n,i}} [Y_{n,i}(1) - \mu_{n,Y(1)}] \end{pmatrix} \\
&= \frac{1}{N_n} \sum_{i=1}^{N_n} \begin{pmatrix} \frac{(1 - D_{n,i}) - (1 - \pi_{n,i})}{1 - \pi_{n,i}} [Y_{n,i}(0) - \mu_{n,Y(0)}] \\ \frac{D_{n,i} - \pi_{n,i}}{\pi_{n,i}} [Y_{n,i}(1) - \mu_{n,Y(1)}] \end{pmatrix} + \frac{1}{N_n} \sum_{i=1}^{N_n} \begin{pmatrix} \frac{(1 - \pi_{n,i})}{1 - \pi_{n,i}} [Y_{n,i}(0) - \mu_{n,Y(0)}] \\ \frac{\pi_{n,i}}{\pi_{n,i}} [Y_{n,i}(1) - \mu_{n,Y(1)}] \end{pmatrix} \\
&= \frac{1}{N_n} \sum_{i=1}^{N_n} \begin{pmatrix} -\frac{e_{n,i}(0)}{1 - \pi_{n,i}} \\ \frac{e_{n,i}(1)}{\pi_{n,i}} \end{pmatrix} [D_{n,i} - \pi_{n,i}] + 0
\end{aligned}$$

Then use this to rewrite the full recentered effect estimator as follows:

$$\begin{aligned}
\sqrt{N_n} V_n^{-1/2} (\hat{\alpha}_{n,ATE} - \alpha_{n,ATE}) &= \sqrt{N_n} V_n^{-1/2} \begin{pmatrix} -1 \\ 1 \end{pmatrix}' \left[\frac{1}{N_n} \sum_{i=1}^{N_n} \begin{pmatrix} -\frac{e_{n,i}(0)}{1 - \pi_{n,i}} \\ \frac{e_{n,i}(1)}{\pi_{n,i}} \end{pmatrix} [D_{n,i} - \pi_{n,i}] \right] (1 + o_p(1)) \\
&= V_n^{-1/2} \left\{ \frac{1}{\sqrt{N_n}} \sum_{i=1}^{N_n} \left[\frac{e_{n,i}(0)}{1 - \pi_{n,i}} + \frac{e_{n,i}(1)}{\pi_{n,i}} \right] [D_{n,i} - \pi_{n,i}] \right\} (1 + o_p(1))
\end{aligned}$$

Notice that this is the quantity already shown to be asymptotically normal, multiplied by $(1 + o_p(1))$. Hence,

$$\sqrt{N_n} V_n^{-1/2} (\hat{\alpha}_{n,ATE} - \alpha_{n,ATE}) \xrightarrow{d} \mathcal{N}(0, 1)$$

A.3.5 Variance Estimation

Proof. (Proposition 5, Limit of $\widehat{V}_{n,AS}$) Recall from Definition 8 that the estimator of Aronow et al. (2017) equals

$$\begin{aligned}\widehat{V}_{n,AS} &= \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{\pi_{n,i}\pi_{n,j}} - \frac{1\{i \neq j\}}{\pi_{n,ij}} \right] (Y_{n,i} - \widehat{\mu}_{n,Y(1)})(Y_{n,j} - \widehat{\mu}_{n,Y(1)})D_{n,i}D_{n,j} \\ &+ \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{(1-\pi_{n,i})(1-\pi_{n,j})} - \frac{1\{i \neq j\}}{1-\pi_{n,i}-\pi_{n,j}+\pi_{n,ij}} \right] (Y_{n,i} - \widehat{\mu}_{n,Y(0)})(Y_{n,j} - \widehat{\mu}_{n,Y(0)})(1-D_{n,i})(1-D_{n,j}) \\ &+ 2 \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j \neq i} \frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{\pi_{n,i}(1-\pi_{n,j})(\pi_{n,i} - \pi_{n,ij})} (Y_{n,i} - \widehat{\mu}_{n,Y(1)})(Y_{n,j} - \widehat{\mu}_{n,Y(0)})D_{n,i}(1-D_{n,j})\end{aligned}$$

Useful for the proof, the infeasible Aronow and Samii estimator replaces $(Y_{n,i} - \widehat{\mu}_{n,Y(d)})$ with $e_{n,i}(d)$ leaving

$$\begin{aligned}\widehat{V}_{n,AS,I} &= \frac{1}{N_n} [A_I + B_I + C_I] \\ A_I &= \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{\pi_{n,i}\pi_{n,j}} - \frac{1\{i \neq j\}}{\pi_{n,ij}} \right] e_{n,i}(1)e_{n,j}(1)D_{n,i}D_{n,j} \\ B_I &= \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{(1-\pi_{n,i})(1-\pi_{n,j})} - \frac{1\{i \neq j\}}{1-\pi_{n,i}-\pi_{n,j}+\pi_{n,ij}} \right] e_{n,i}(0)e_{n,j}(0)(1-D_{n,i})(1-D_{n,j}) \\ C_I &= 2 \sum_{i=1}^{N_n} \sum_{j \neq i} \frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{\pi_{n,i}(1-\pi_{n,j})(\pi_{n,i} - \pi_{n,ij})} e_{n,i}(1)e_{n,j}(0)(D_{n,i} - D_{n,i}D_{n,j})\end{aligned}$$

From here, the proof strategy will be broken up into steps as follows:

- (a) Derive expressions for $\widehat{V}_{n,AS,I}$, $\mathbb{E}[\widehat{V}_{n,AS,I}]$, and the difference $\widehat{V}_{n,AS,I} - \mathbb{E}[\widehat{V}_{n,AS,I}]$
- (b) Rewrite $\widehat{V}_{n,AS,I} - \mathbb{E}[\widehat{V}_{n,AS,I}]$ (a double sum) as $\frac{1}{N_n} \sum_{i=1}^{N_n} W_{n,i} + R$, where each $W_{n,i}$ is a sum involving a limited number of units centered around location i and cross terms between i and all remaining units are directed into R , for all i .
- (c) Break up $\text{Var}(\sum_{i=1}^{N_n} W_{n,i}) = \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \text{Cov}(W_{n,i}, W_{n,j})$ into two pieces: a double sum where i and j are “near” and another double sum where they are “far”. Derive a bound on the “near” observations using Cauchy-Schwarz (which assumes perfect correlation but does not matter since the number of near observations will be limited) and a bound on the “far” observations using the covariance inequality of Lemma 8 (which will make use of the mixing assumption and decay in dependence).
- (d) Bound moments of $W_{n,i}$.
- (e) Put together the results of Step (c) and (d).

(f) Bound the remainder term R in Step (b).

(g) Show convergence of infeasible estimator, $\widehat{V}_{n,AS,I} - \mathbb{E}[\widehat{V}_{n,AS,I}] \xrightarrow{P} 0$ using Chebyshev's inequality and the results from Steps (e) and (f).

(h) Relating the mean of the estimator to true variance plus “conservativeness”, i.e. $\mathbb{E}[\widehat{V}_{n,AS,I}] = V_n + \frac{1}{N_n} \sum_{i=1}^{N_n} (e_{n,i}(1) - e_{n,i}(0))^2$

(i) Feasible and Infeasible equivalent: Show that $\widehat{V}_{n,AS,I}$ —which used the unknown $e_{n,i}(d) = Y_{n,i}(d) - \mu_{n,Y(d)}$ —is asymptotically equivalent to using the feasible alternative $V_{n,AS}$ —which used $Y_{n,i} - \widehat{\mu}_{n,Y(d)}$, i.e. $\widehat{V}_{n,AS} - \mathbb{E}[\widehat{V}_{n,AS,I}] \xrightarrow{P} 0$:

(Step(a)) It is convenient to rewrite A_I and B_I , removing $\mathbf{1}\{i \neq j\}$ terms. First, rewrite A_I

$$\begin{aligned} A_I &= \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{\pi_{n,i}\pi_{n,j}} - \frac{1}{\pi_{n,ij}} \right] e_{n,i}(1)e_{n,j}(1)D_{n,i}D_{n,j} + \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{\pi_{n,ij}} - \frac{\mathbf{1}\{i \neq j\}}{\pi_{n,ij}} \right] e_{n,i}(1)e_{n,j}(1)D_{n,i}D_{n,j} \\ &= \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{\pi_{n,i}\pi_{n,j}} - \frac{1}{\pi_{n,ij}} \right] e_{n,i}(1)e_{n,j}(1)D_{n,i}D_{n,j} + \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1 - \mathbf{1}\{i \neq j\}}{\pi_{n,ij}} \right] e_{n,i}(1)e_{n,j}(1)D_{n,i}D_{n,j} \\ &= \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{\pi_{n,ij}} \right] \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} D_{n,i}D_{n,j} + \sum_{i=1}^{N_n} \frac{e_{n,i}(1)^2}{\pi_{n,i}} D_{n,i} \end{aligned}$$

Similarly, B_I can be rewritten

$$B_I = \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{1 - \pi_{n,i} - \pi_{n,j} + \pi_{n,ij}} \right] \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \frac{e_{n,j}(0)}{1 - \pi_{n,j}} [1 - D_{n,i} - D_{n,j} + D_{n,i}D_{n,j}] + \sum_{i=1}^{N_n} \frac{e_{n,i}(0)^2}{1 - \pi_{n,i}} (1 - D_{n,i})$$

Putting everything together, these expressions imply that

$$N_n \cdot (\widehat{V}_{n,AS,I} - \mathbb{E}[\widehat{V}_{n,AS,I}]) = \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} X_{n,ij}$$

where, after some manipulation,

$$\begin{aligned}
X_{n,ij} = & \left\{ \left[\frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{\pi_{n,ij}} \right] \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} + \left[\frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{1 - \pi_{n,i} - \pi_{n,j} + \pi_{n,ij}} \right] \frac{e_{n,i}(0)}{(1 - \pi_{n,i})} \frac{e_{n,j}(0)}{(1 - \pi_{n,j})} \right. \\
& \left. - 2 \left[\frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{\pi_{n,i} - \pi_{n,ij}} \right] \mathbf{1}\{i \neq j\} \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(0)}{(1 - \pi_{n,j})} \right\} \cdot [D_{n,i}D_{n,j} - \pi_{n,ij}] \\
& + \left\{ \left[\frac{1}{\pi_{n,i}} e_{n,i}(1)^2 - \frac{1}{1 - \pi_{n,i}} e_{n,i}(0)^2 \right] \mathbf{1}\{i = j\} + \left[2 \frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{(\pi_{n,i} - \pi_{n,ij})} \right] \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(0)}{(1 - \pi_{n,j})} \mathbf{1}\{i \neq j\} \right. \\
& \left. - 2 \left[\frac{\pi_{n,ij} - \pi_{n,j}\pi_{n,i}}{1 - \pi_{n,j} - \pi_{n,i} + \pi_{n,ij}} \right] \frac{e_{n,i}(0)}{(1 - \pi_{n,j})} \frac{e_{n,j}(0)}{(1 - \pi_{n,i})} \right\} \cdot [D_{n,i} - \pi_{n,i}]
\end{aligned}$$

(Step (b)) The double sum expression above for $N_n \cdot (\widehat{V}_{n,AS,I} - \mathbb{E}[\widehat{V}_{n,AS,I}])$ walks over all pairs i and j , but it is useful to recognize that far apart i and j will imply small $X_{n,ij}$. This is because far apart units will be nearly independent and so $\text{Cov}(D_{n,i}, D_{n,j}) = \pi_{n,ij} - \pi_{n,i}\pi_{n,j} \approx 0$, which implies a small $X_{n,ij}$. In light of this fact, break up the full expression as follows, for some sequence of cutoffs b_n (that we will select later):

$$N_n \cdot (\widehat{V}_{n,AS,I} - \mathbb{E}[\widehat{V}_{n,AS,I}]) = \underbrace{\sum_{i=1}^{N_n} \left(\sum_{j: d(\ell_i, \ell_j) \leq b_n} X_{n,ij} \right)}_{W_{n,i}} + \underbrace{\sum_{i=1}^{N_n} \left(\sum_{j: d(\ell_i, \ell_j) > b_n} X_{n,ij} \right)}_{N_n \cdot R(b_n)} \quad (\text{A.4})$$

Notice that the first double sum is like a kernel variance estimator with truncation up to distance b_n . Therefore, proving that this is convergent will follow the same approach as for a kernel estimator. The second portion is a “remainder” term that we already argued will be small, and which I will show formally. Together, convergence in probability will follow.

(Step (c)) This step in which I characterize $\text{Var}(\sum_{i=1}^{N_n} W_{n,i})$ largely follows [Conley \(1999\)](#) and other proofs of kernel estimator consistency. To give an overview, note that I can write

$$\text{Var} \left(\sum_{i=1}^{N_n} W_{n,i} \right) = \left| \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \mathbb{E}[W_{n,i} W_{n,k}] \right| \leq \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |\mathbb{E}[W_{n,i} W_{n,k}]|$$

The latter expression is a double sum of cross terms, where each $W_{n,i}$ is a “ball” of units centered around corresponding ℓ_i . Because of weak dependence, far apart $W_{n,i}$ will be near independent, and if the rate of growth of the $W_{n,i}$ is controlled as well, the entire variance expression can be controlled and Chebyshev’s inequality can be used to establish convergence.

Therefore, to start, separate the double sum into “near” and “far” observations:

$$\text{Var} \left(\sum_{i=1}^{N_n} W_{n,i} \right) \leq \sum_{i=1}^{N_n} \sum_{d(\ell_i, \ell_k) \leq 4b_n} |\mathbb{E}[W_{n,i} W_{n,k}]| + \sum_{i=1}^{N_n} \sum_{d(\ell_i, \ell_k) > 4b_n} |\mathbb{E}[W_{n,i} W_{n,k}]| \quad (\text{A.5})$$

The cutoff $4b_n$ is chosen to ensure we can easily bound the second double sum. To see why, recall that each $W_{n,i}$ is a sum over $D_{n,k}$ within distance b_n of ℓ_i . Separation by $4b_n$ will ensure that the collections of units around ℓ_i and ℓ_k are “sufficiently separated” so that the second double sum can be controlled by the weak dependence assumptions on the $D_{n,i}$.

For the first double sum in Expression A.5, I will use Cauchy-Schwarz. Although this is the worst possible (i.e. perfect correlation) bound, it is for only a subset of terms in the variance expression. Therefore, I start with

$$|\mathbb{E}[W_{n,i} W_{n,k}]| \leq \|W_{n,i}\|_2 \cdot \|W_{n,k}\|_2 \leq \sup_{\ell_i \in \mathcal{I}_n} \|W_{n,i}\|_2^2$$

Using this, the first double sum in Expression A.5 satisfies the bound

$$\sum_{i=1}^{N_n} \sum_{d(\ell_i, \ell_k) \leq 4b_n} |\mathbb{E}[W_{n,i} W_{n,k}]| \leq \sum_{i=1}^{N_n} \sum_{d(\ell_i, \ell_k) \leq 4b_n} \sup_{\ell_j \in \mathcal{I}_n} \|W_{n,j}\|_2^2 = N_n \frac{8(4b_n)^2}{d_{\min}^2} \cdot \sup_{\ell_j \in \mathcal{I}_n} \|W_{n,j}\|_2^2$$

where the counts of observations followed from Assumption 4 and Lemma 9. This handles the “near” observations. (An expression for $\sup_j \|W_{n,j}\|_2^2$ is derived in Step (d) below.)

The next task is bounding the “far” observations, as contained in the second double sum in Expression A.5. For that, I will use the covariance inequality of Lemma 8 that follows from the assumed mixing conditions. That is justified for two reasons. First, $W_{n,i}$ and $W_{n,k}$ in the second sum are separated by more than distance $4b_n$, and so no $D_{n,a}$ in the sum of units comprising $W_{n,i}$ are within $2b_n$ of a $D_{n,b}$ in $W_{n,k}$.¹ Second, as we’ll see in Step (d), for any fixed n and finite $\delta > 0$, $\|W_{n,i}\|_{2+\delta}$ is finite. So invoke Lemma 8 to write

$$|\mathbb{E}[W_{n,\ell_i} W_{n,\ell_j}]| \leq 8 \cdot \alpha_{\infty, \infty}(2b_n)^{\frac{\delta}{2+\delta}} \sup_{\ell_k \in \mathcal{I}_n} \|W_{n,k}\|_{2+\delta}^2 \leq 8 \cdot \alpha_{\infty, \infty}(b_n)^{\frac{\delta}{2+\delta}} \sup_{\ell_k \in \mathcal{I}_n} \|W_{n,k}\|_{2+\delta}^2$$

¹This is how I go from working with $W_{n,i}$ to an inequality based on mixing coefficients of the $D_{n,i}$

The second sum in Expression A.5 is then bounded as follows:

$$\begin{aligned} \sum_{i=1}^{N_n} \sum_{d(\ell_i, \ell_k) > 4b_n} |\mathbb{E}[W_{n,i} W_{n,k}]| &\leq \sum_{i=1}^{N_n} \sum_{d(\ell_i, \ell_k) > 4b_n} 8 \cdot \alpha_{\infty, \infty}(b_n)^{\frac{\delta}{2+\delta}} \sup_{\ell_j \in \mathcal{I}_n} \|W_{n,j}\|_{2+\delta}^2 \\ &\leq N_n^2 \cdot 8 \cdot \alpha_{\infty, \infty}(b_n)^{\frac{\delta}{2+\delta}} \sup_{\ell_j \in \mathcal{I}_n} \|W_{n,j}\|_{2+\delta}^2 \end{aligned}$$

The final inequality followed because the double sum has at most N_n^2 units.

Putting together both pieces, Expression A.5 becomes

$$\text{Var} \left(\sum_{i=1}^{N_n} W_{n,i} \right) \leq \frac{32}{d_{\min}^2} \cdot N_n b_n^2 \sup_{\ell_j \in \mathcal{I}_n} \|W_{n,j}\|_2^2 + 8N_n^2 \alpha_{\infty, \infty}(b_n)^{\frac{\delta}{2+\delta}} \sup_{\ell_j \in \mathcal{I}_n} \|W_{n,j}\|_{2+\delta}^2$$

(Step (d)) To further characterize the bounds in the previous step, I now bound moments of $W_{n,i}$. Specifically, Minkowski's inequality for any $q > 2$ implies

$$\begin{aligned} \|W_{n,i}\|_q &\leq \sum_{j: d(\ell_i, \ell_j) \leq b_n} \left| \left\{ \left[\frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{\pi_{n,ij}} \right] \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} + \left[\frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{1 - \pi_{n,i} - \pi_{n,j} + \pi_{n,ij}} \right] \frac{e_{n,i}(0)}{(1 - \pi_{n,i})} \frac{e_{n,j}(0)}{(1 - \pi_{n,j})} \right. \right. \\ &\quad \left. \left. - 2 \left[\frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{\pi_{n,i} - \pi_{n,ij}} \right] \mathbf{1}\{i \neq j\} \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(0)}{(1 - \pi_{n,j})} \right\} \right| \cdot \|D_{n,i} D_{n,j} - \pi_{n,ij}\|_q \\ &+ \sum_{j: d(\ell_i, \ell_j) \leq b_n} \left| \left\{ \left[\frac{1}{\pi_{n,i}} e_{n,i}(1)^2 - \frac{1}{1 - \pi_{n,i}} e_{n,i}(0)^2 \right] \mathbf{1}\{i = j\} + \left[2 \frac{\pi_{n,ij} - \pi_{n,i}\pi_{n,j}}{(\pi_{n,i} - \pi_{n,ij})} \right] \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(0)}{(1 - \pi_{n,j})} \mathbf{1}\{i \neq j\} \right. \right. \\ &\quad \left. \left. - 2 \left[\frac{\pi_{n,ij} - \pi_{n,j}\pi_{n,i}}{1 - \pi_{n,j} - \pi_{n,i} + \pi_{n,ij}} \right] \frac{e_{n,i}(0)}{(1 - \pi_{n,j})} \frac{e_{n,j}(0)}{(1 - \pi_{n,i})} \right\} \right| \cdot \|D_{n,i} - \pi_{n,i}\|_q \end{aligned}$$

By bounded potential outcomes from [Assumption 1](#) (translating into bounded $e_{n,i}(d)$), nondegenerate treatment probabilities from [Assumption 2](#), and $D_{n,i} \in \{0, 1\}$, it is possible to find bounds for terms in the expression above that do not depend upon n (see normality proof for an example of how these assumptions were used to establish such a bound), which leads to

$$\begin{aligned} \|W_{n,i}\|_q &\leq \sum_{j: d(\ell_i, \ell_j) \leq b_n} c_{1,q} \|D_{n,i} D_{n,j} - \pi_{n,ij}\|_q + \sum_{j: d(\ell_i, \ell_j) \leq b_n} c_{2,q} \|D_{n,i} - \pi_{n,i}\|_q \\ &\leq \sum_{j: d(\ell_i, \ell_j) \leq b_n} (c_{1,q} + c_{2,q}) \\ &\leq (c_{1,q} + c_{2,q}) \cdot 8(b_n/d_{\min})^2 = c_{3,q} b_n^2 \end{aligned}$$

where the bound and count on the number of units in the ball followed from Assumption 4 and Lemma 9, and where the $c_{1,q}, c_{2,q}, c_{3,q}$ are constants that do not depend upon n .

(Step (e)) Plug the bounds from Step (d) into the final inequality in Step (c) to write

$$\begin{aligned} \text{Var} \left(\sum_{i=1}^{N_n} W_{n,i} \right) &\leq \frac{32}{d_{\min}^2} N_n b_n^2 \cdot (c_{3,2} b_n^2)^2 + 8 N_n^2 \cdot \alpha_{\infty, \infty}(b_n)^{\frac{\delta}{2+\delta}} (c_{3,2+\delta} b_n^2)^2 \\ &= c_4 N_n b_n^6 + c_5 \alpha_{\infty, \infty}(b_n)^{\frac{\delta}{2+\delta}} N_n^2 b_n^4 \end{aligned}$$

where c_4 and c_5 are again constants that do not depend upon n .

(Step (f)) Consider i and j such that $d(\ell_i, \ell_j) > b_n$, as in the definition of the remainder from Step (b). Because of bounded potential outcomes from Assumption 1 (translating into bounded $e_{n,i}(d)$), nondegenerate treatment probabilities from Assumption 2, and $D_{n,i} \in \{0, 1\}$, it is possible to find a constant C that does not depend upon n (see normality proof for an example of how these assumptions were used to establish such a bound) such that

$$|X_{n,ij}| \leq C |\pi_{n,ij} - \pi_{n,i} \pi_{n,j}|$$

Next, by Lemma 8 and the fact that $\alpha_{1,1}(m) \leq \alpha_{\infty, \infty}(m)$,

$$|X_{n,ij}| \leq 8C \cdot \alpha_{1,1}(d(\ell_i, \ell_j))^{\frac{\delta}{2+\delta}} \leq 8C \cdot \alpha_{1,1}(b_n)^{\frac{\delta}{2+\delta}} \leq 8C \cdot \alpha_{\infty, \infty}(b_n)^{\frac{\delta}{2+\delta}}$$

This implies for the remainder term that

$$|R(b_n)| \leq \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j : d(\ell_i, \ell_j) > b_n} |X_{n,ij}| \leq \frac{1}{N_n} \cdot N_n^2 \cdot 8C \cdot \alpha_{\infty, \infty}(b_n)^{\frac{\delta}{2+\delta}} = N_n^{-1} 8C \cdot \alpha_{\infty, \infty}(b_n)^{\frac{\delta}{2+\delta}}$$

where the inequality followed from the bound derived on $|X_{n,ij}|$ and the fact that the double sum has at most N_n^2 units. Below, I'll show that choice of bandwidth and the mixing assumptions will imply that this is $o(1)$.

(Step (g)) In this step, I show that the variance goes to zero, implying convergence in probability. To

do so, first note that Expression A.4 implies

$$|\widehat{V}_{n,AS,I} - \mathbb{E}[\widehat{V}_{n,AS,I}]| \leq \left| \frac{1}{N_n} \sum_{i=1}^{N_n} W_{n,i} \right| + |R(b_n)|$$

This implies that

$$P \left[|\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}]| \geq \nu \right] \leq P \left[\left| \frac{1}{N_n} \sum_{i=1}^{N_n} W_{n,i} \right| + |R(b_n)| \geq \nu \right]$$

Rearrange and apply Chebyshev's inequality for any fixed ν and n :

$$\begin{aligned} P \left[|\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}]| \geq \nu \right] &\leq P \left[\left| \frac{1}{N_n} \sum_{i=1}^{N_n} W_{n,i} \right| \geq \nu - |R(b_n)| \right] \\ &\leq \frac{1}{(\nu - |R(b_n)|)^2} \text{Var} \left(\frac{1}{N_n} \sum_{i=1}^{N_n} W_{n,i} \right) \end{aligned}$$

Use the bound on the variance from Step (e), which implies

$$P \left[|\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}]| \geq \nu \right] \leq \frac{1}{(\nu - |R(b_n)|)^2} \frac{1}{N_n^2} [c_4 N_n b_n^6 + c_5 \alpha_{\infty, \infty}(b_n)^{\frac{\delta}{2+\delta}} N_n^2 b_n^4]$$

The above bound applies for any choice of b_n , and so it is a family of bounds. To derive a “tight enough” bound, I choose a particular sequence under which the righthand side expression tends toward zero.

First, note that by [Assumption 7](#), for some $\varepsilon, k > 0$, $\alpha_{\infty, \infty}(m)^{\frac{\delta}{2+\delta}} \leq k m^{-(4+\varepsilon)}$. Therefore,

$$\begin{aligned} P \left[|\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}]| \geq \nu \right] &\leq \frac{1}{(\nu - |R(b_n)|)^2} \frac{1}{N_n^2} [c_4 N_n b_n^6 + c_5 k b_n^{-(4+\varepsilon)} N_n^2 b_n^4] \\ &\leq \frac{1}{(\nu - |R(b_n)|)^2} [c_4 N_n^{-1} b_n^6 + c_5 \tilde{k} b_n^{-\varepsilon}] \end{aligned}$$

as well as

$$|R(b_n)| = N_n^{-1} \tilde{C} b_n^{-(4+\varepsilon)}$$

Now choose $b_n = N_n^{\frac{1-\eta}{6}}$ for some $\eta \in (0, 1)$. Plugging in

$$P \left[|\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}]| \geq \nu \right] \leq \frac{1}{(\nu - |R(b_n)|)^2} [c_4 N_n^{-\eta} + c_5 \tilde{k}_2 N_n^{-\varepsilon \frac{1-\eta}{6}}]$$

as well as

$$|R(b_n)| = \tilde{C} N_n^{-\frac{(1-\eta)(4+\varepsilon)}{6}-1} = o(1)$$

Because N_n is increasing, by the expressions above, for any fixed ν ,

$$P \left[|\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}]| \geq \nu \right] \leq \frac{1}{(v - |R(b_n)|)^2} [c_4 N_n^{-\eta} + c_5 \tilde{k}_2 N_n^{-\varepsilon \frac{1-\eta}{6}}] \rightarrow 0$$

Because ν was arbitrary, this implies $\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}] \xrightarrow{p} 0$.

(Step (h)) Given the expressions derived above, the expectation of the infeasible estimator is

$$\begin{aligned} N_n \cdot \mathbb{E}[\widehat{V}_{n,AS,I}] &= \mathbb{E}[A_I] + \mathbb{E}[B_I] + \mathbb{E}[C_I] \\ &= \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} [\pi_{n,ij} - \pi_{n,i}\pi_{n,j}] \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} + \sum_{i=1}^{N_n} e_{n,i}(1)^2 \\ &\quad + \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} [\pi_{n,ij} - \pi_{n,i}\pi_{n,j}] \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \frac{e_{n,j}(0)}{1 - \pi_{n,j}} + \sum_{i=1}^{N_n} e_{n,i}(0)^2 \\ &\quad + 2 \sum_{i=1}^{N_n} \sum_{j \neq i} [\pi_{n,ij} - \pi_{n,i}\pi_{n,j}] \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(0)}{1 - \pi_{n,j}} \end{aligned}$$

Simplify this expression to

$$\begin{aligned} \mathbb{E}[\widehat{V}_{n,AS,I}] &= \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left(\frac{e_{n,i}(1)}{\pi_{n,i}} + \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \right) \left(\frac{e_{n,j}(1)}{\pi_{n,j}} + \frac{e_{n,j}(0)}{1 - \pi_{n,j}} \right) [\pi_{n,ij} - \pi_{n,i}\pi_{n,j}] \\ &\quad + \frac{1}{N_n} \sum_{i=1}^{N_n} [e_{n,i}(1) - e_{n,i}(0)]^2 \\ &= V_n + \frac{1}{N_n} \sum_{i=1}^{N_n} [e_{n,i}(1) - e_{n,i}(0)]^2 \end{aligned}$$

This and the result from Step (g) together imply

$$\widehat{V}_{n,AS,I} - V_n - \frac{1}{N_n} \sum_{i=1}^{N_n} [e_{n,i}(1) - e_{n,i}(0)]^2 \xrightarrow{p} 0$$

(Step (i)) Finally, to show that the infeasible and feasible estimators have equivalent probability limits, I

consider one piece of the variance estimator, A_I and A . Comparing the infeasible and feasible:

$$A_I = \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{\pi_{n,i}\pi_{n,j}} - \frac{1\{i \neq j\}}{\pi_{n,ij}} \right] (Y_{n,i}(1) - \mu_{n,Y(1)})(Y_{n,j}(1) - \mu_{n,Y(1)})D_{n,i}D_{n,j}$$

$$A = \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{\pi_{n,i}\pi_{n,j}} - \frac{1\{i \neq j\}}{\pi_{n,ij}} \right] (Y_{n,i}(1) - \hat{\mu}_{n,Y(1)})(Y_{n,j}(1) - \hat{\mu}_{n,Y(1)})D_{n,i}D_{n,j}$$

Take difference and simplify to

$$A_I - A = \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{1}{\pi_{n,i}\pi_{n,j}} - \frac{1\{i \neq j\}}{\pi_{n,ij}} \right] [2Y_{n,i}(1)(\hat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}) + \mu_{n,Y(1)}^2 - \hat{\mu}_{n,Y(1)}^2] D_{n,i}D_{n,j}$$

This can be rewritten as

$$A_I - A = \sum_{i=1}^{N_n} [2Y_{n,i}(1)(\hat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}) + \mu_{n,Y(1)}^2 - \hat{\mu}_{n,Y(1)}^2] \frac{D_{n,i}}{\pi_{n,i}}$$

$$+ \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} [\pi_{n,ij} - \pi_{n,i}\pi_{n,j}] \frac{1}{\pi_{n,ij}\pi_{n,i}\pi_{n,j}} [2Y_{n,i}(1)(\hat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}) + \mu_{n,Y(1)}^2 - \hat{\mu}_{n,Y(1)}^2] D_{n,i}D_{n,j}$$

Using the triangle inequality, pulling out terms, and dividing by N_n gives

$$\frac{1}{N_n} |A_I - A| \leq |\hat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}| \cdot \frac{1}{N_n} \sum_{i=1}^{N_n} |2Y_{n,i}(1)| \cdot \frac{D_{n,i}}{\pi_{n,i}}$$

$$+ |\mu_{n,Y(1)}^2 - \hat{\mu}_{n,Y(1)}^2| \cdot \frac{1}{N_n} \sum_{i=1}^{N_n} \frac{D_{n,i}}{\pi_{n,i}}$$

$$+ |\hat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}| \cdot \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |\pi_{n,ij} - \pi_{n,i}\pi_{n,j}| \frac{2|Y_{n,i}(1)|D_{n,i}D_{n,j}}{|\pi_{n,ij}\pi_{n,i}\pi_{n,j}|}$$

$$+ |\mu_{n,Y(1)}^2 - \hat{\mu}_{n,Y(1)}^2| \cdot \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |\pi_{n,ij} - \pi_{n,i}\pi_{n,j}| \frac{D_{n,i}D_{n,j}}{|\pi_{n,ij}\pi_{n,i}\pi_{n,j}|}$$

Because of bounded potential outcomes from [Assumption 1](#), nondegenerate treatment probabilities from [Assumption 2](#), and $D_{n,i} \in \{0, 1\}$, it is possible to find bounds K_1, K_2, K_3, K_4 on the terms within the sums

and rewrite

$$\begin{aligned}
\frac{1}{N_n} |A_I - A| &\leq |\widehat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}| \cdot K_1 + |\mu_{n,Y(1)}^2 - \widehat{\mu}_{n,Y(1)}^2| \cdot K_2 \\
&\quad + |\widehat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}| \cdot \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |\pi_{n,ij} - \pi_{n,i}\pi_{n,j}| K_3 \\
&\quad + |\mu_{n,Y(1)}^2 - \widehat{\mu}_{n,Y(1)}^2| \cdot \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} |\pi_{n,ij} - \pi_{n,i}\pi_{n,j}| K_4
\end{aligned}$$

Next, by Lemma 10, the double sum is $O(1)$, hence

$$\frac{1}{N_n} |A_I - A| \leq |\widehat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}| \cdot (K_1 + O(1) \cdot K_3) + |\mu_{n,Y(1)}^2 - \widehat{\mu}_{n,Y(1)}^2| \cdot (K_2 + O(1) \cdot K_4)$$

By consistency from Proposition 2, the terms converge to zero as $n \rightarrow \infty$. Exactly the same follows for B_I and C_I . $\square$

Proof. (Proposition 6, Limit of $\widehat{V}_{n,C}$) Recall $X_{n,i} = (1 - D_{n,i}, D_{n,i})'$ and

$$\omega_{n,i} = \frac{D_{n,i}}{\pi_{n,i}} + \frac{1 - D_{n,i}}{1 - \pi_{n,i}}.$$

The Conley variance estimator for a given kernel equals

$$\begin{aligned}
\widehat{V}_{n,C} &= N_n \begin{pmatrix} -1 & 1 \end{pmatrix} (X_n' \omega_n X_n)^{-1} \left[\sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \omega_{n,i} \omega_{n,j} \widehat{e}_{n,i} \widehat{e}_{n,j} X_{n,i} X_{n,j}' K_{n,ij} \right] (X_n' \omega_n X_n)^{-1} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \\
&= \begin{pmatrix} -1 & 1 \end{pmatrix} (X_n \omega_n X_n / N_n)^{-1} \left[\frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \widehat{e}_{n,i} \widehat{e}_{n,j} \begin{pmatrix} \frac{1-D_{n,i}}{1-\pi_{n,i}} & \frac{1-D_{n,j}}{1-\pi_{n,j}} & \frac{1-D_{n,i}}{1-\pi_{n,i}} \frac{D_{n,j}}{\pi_{n,j}} \\ \frac{D_{n,i}}{\pi_{n,i}} & \frac{1-D_{n,j}}{1-\pi_{n,j}} & \frac{D_{n,i}}{\pi_{n,i}} \frac{D_{n,j}}{\pi_{n,j}} \end{pmatrix} K_{n,ij} \right] (X_n \omega_n X_n / N_n)^{-1} \begin{pmatrix} -1 \\ 1 \end{pmatrix}
\end{aligned}$$

Using the same approach that was used to prove Proposition 2, can show that the “bread” $(X_n \omega_n X_n / N_n)^{-1}$ in the sandwich estimator converges in probability to I_n . Therefore, write

$$\widehat{V}_{n,C} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}' \left[\frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \widehat{e}_{n,i} \widehat{e}_{n,j} \begin{pmatrix} \frac{1-D_{n,i}}{1-\pi_{n,i}} & \frac{1-D_{n,j}}{1-\pi_{n,j}} & \frac{1-D_{n,i}}{1-\pi_{n,i}} \frac{D_{n,j}}{\pi_{n,j}} \\ \frac{D_{n,i}}{\pi_{n,i}} & \frac{1-D_{n,j}}{1-\pi_{n,j}} & \frac{D_{n,i}}{\pi_{n,i}} \frac{D_{n,j}}{\pi_{n,j}} \end{pmatrix} K_{n,ij} \right] \begin{pmatrix} -1 \\ 1 \end{pmatrix} + o_P(1)$$

The central task therefore lies in proving convergence of the “meat.” This task will be broken up into subtasks that make use also of the “infeasible” Conley estimator that replaces

$$\widehat{e}_{n,i} = Y_{n,i} - \widehat{\mu}_{n,1} D_{n,i} - \widehat{\mu}_{n,0} (1 - D_{n,i})$$

with $e_{n,i}$. This infeasible estimator is defined

$$\hat{V}_{n,C,I} := \begin{pmatrix} -1 \\ 1 \end{pmatrix}' \left[\frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} e_{n,i} e_{n,j} \begin{pmatrix} \frac{1-D_{n,i}}{1-\pi_{n,i}} \frac{1-D_{n,j}}{1-\pi_{n,j}} & \frac{1-D_{n,i}}{1-\pi_{n,i}} \frac{D_{n,j}}{\pi_{n,j}} \\ \frac{D_{n,i}}{\pi_{n,i}} \frac{1-D_{n,j}}{1-\pi_{n,j}} & \frac{D_{n,i}}{\pi_{n,i}} \frac{D_{n,j}}{\pi_{n,j}} \end{pmatrix} K_{n,ij} \right] \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

From here, the proof strategy will be as follows (and similar to the proof of Proposition 5):

- (a) Derive expressions for $\hat{V}_{n,C,I}$, $\mathbb{E}[\hat{V}_{n,C,I}]$, and the difference $\hat{V}_{n,C,I} - \mathbb{E}[\hat{V}_{n,C,I}]$
 - (b) Rewrite $\hat{V}_{n,C,I} - \mathbb{E}[\hat{V}_{n,C,I}]$ (a double sum) as $\frac{1}{N_n} \sum_{i=1}^{N_n} W_{n,i}$, where each $W_{n,i}$ a sum of units centered around location i .
 - (c) Break up $\text{Var}(\sum_{i=1}^{N_n} W_{n,i}) = \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \text{Cov}(W_{n,i}, W_{n,j})$ into two pieces: a double sum where i and j are “near” and another where they are “far”. Derive a bound on the “near” observations using Cauchy-Schwarz (which assumes perfect correlation but does not matter since the number of near observations will be limited) and a bound on the “far” observations using the covariance inequality of Lemma 8 (which will make use of the mixing assumption and decay in dependence).
 - (d) Bound moments of $W_{n,i}$.
 - (e) Put together the results of Step (c) and (d).
 - (f) Show convergence of infeasible estimator, $\hat{V}_{n,C,I} - \mathbb{E}[\hat{V}_{n,C,I}] \xrightarrow{P} 0$ using Chebyshev’s inequality and the result from Steps (e).
 - (g) Feasible and Infeasible equivalent: Show that replacing $e_{n,i}(d)$ with $\hat{e}_{n,i}(d)$ makes no difference asymptotically and so $\hat{V}_{n,C} - \mathbb{E}[\hat{V}_{n,C,I}] \xrightarrow{P} 0$:
 - (h) Relating $\mathbb{E}[\hat{V}_{n,C,I}]$ to the true variance V_n .
- (Step (a)) As a first step for the infeasible estimator, note that

$$\begin{aligned} e_{n,i} &= Y_{n,i} - \mu_{n,Y(1)} D_{n,i} - \mu_{n,Y(0)} (1 - D_{n,i}) \\ &= [Y_{n,i}(1) - \mu_{n,Y(1)}] D_{n,i} + [Y_{n,i}(0) - \mu_{n,Y(0)}] (1 - D_{n,i}) \\ &= e_{n,i}(1) D_{n,i} + e_{n,i}(0) (1 - D_{n,i}) \end{aligned}$$

Substitute in this expression for $e_{n,i}$ and simplify to

$$\begin{aligned}\widehat{V}_{n,C,I} &= \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} D_{n,i} D_{n,j} K_{n,ij} + \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(0)}{1-\pi_{n,i}} \frac{e_{n,j}(0)}{1-\pi_{n,j}} (1-D_{n,i})(1-D_{n,j}) K_{n,ij} \\ &\quad - \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(0)}{1-\pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} (1-D_{n,i}) D_{n,j} K_{n,ij} - \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(0)}{1-\pi_{n,j}} D_{n,i} (1-D_{n,j}) K_{n,ij}\end{aligned}$$

Note that the expectation of this equals

$$\begin{aligned}\mathbb{E}[\widehat{V}_{n,C,I}] &= \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} [\pi_{n,ij}] K_{n,ij} + \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(0)}{1-\pi_{n,i}} \frac{e_{n,j}(0)}{1-\pi_{n,j}} [1-\pi_{n,i}-\pi_{n,j}+\pi_{n,ij}] K_{n,ij} \\ &\quad - \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(0)}{1-\pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} [\pi_{n,j}-\pi_{n,ij}] K_{n,ij} - \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(0)}{1-\pi_{n,j}} [\pi_{n,i}-\pi_{n,ij}] K_{n,ij}\end{aligned}$$

Given these expressions, the difference can be computed and simplified to

$$\begin{aligned}\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}] &= \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left(\frac{e_{n,i}(1)}{\pi_{n,i}} + \frac{e_{n,i}(0)}{1-\pi_{n,i}} \right) \left[\left(\frac{e_{n,j}(1)}{\pi_{n,j}} + \frac{e_{n,j}(0)}{1-\pi_{n,j}} \right) [D_{n,i} D_{n,j} - \pi_{n,ij}] + 2 \frac{e_{n,j}(0)}{1-\pi_{n,j}} (\pi_{n,i} - D_{n,i}) \right] K_{n,ij}\end{aligned}$$

where $W_{n,i}$ is a mean-zero random variable.

(Step (b)) It is possible to rewrite

$$\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}] = \frac{1}{N_n} \sum_{i=1}^{N_n} W_{n,i}$$

where

$$\begin{aligned}W_{n,i} &= \sum_{j: d(\ell_i, \ell_j) \leq b_n} \left(\frac{e_{n,i}(1)}{\pi_{n,i}} + \frac{e_{n,i}(0)}{1-\pi_{n,i}} \right) \left[\left(\frac{e_{n,j}(1)}{\pi_{n,j}} + \frac{e_{n,j}(0)}{1-\pi_{n,j}} \right) [D_{n,i} D_{n,j} - \pi_{n,ij}] \right. \\ &\quad \left. + 2 \frac{e_{n,j}(0)}{1-\pi_{n,j}} (\pi_{n,i} - D_{n,i}) \right] K_{n,ij}\end{aligned}$$

Notice that $W_{n,i}$ only ranges over j such that $d(\ell_i, \ell_j) \leq b_n$. This followed because, by construction (see Definition 9), the kernel implies distance-based truncation where $K_{n,ij} = 0$ if $d(\ell_i, \ell_j) > b_n$ where b_n is a Euclidean radius/distance/bandwidth that grows with n .

(Step (c)) Follow Step (c) of the proof of Proposition 5 exactly, but implicitly with the alternative definition of $W_{n,i}$ from Step (b) in this proof. This gives

$$\text{Var} \left(\sum_{i=1}^{N_n} W_{n,i} \right) \leq \frac{32}{d_{\min}^2} N_n b_n^2 \cdot \sup_{\ell_j \in \mathcal{I}_n} \|W_{n,j}\|_2^2 + 8N_n^2 \alpha_{\infty, \infty}(b_n)^{\frac{\delta}{2+\delta}} \sup_{\ell_j \in \mathcal{I}_n} \|W_{n,j}\|_{2+\delta}^2$$

(Step (d)) To complete the bound, bound the moments of $W_{n,i}$. Specifically, Minkowski's inequality for any $q > 2$ and the $W_{n,i}$ as defined in Step (b),

$$\begin{aligned} \|W_{n,i}\|_q \leq & \sum_{j: d(\ell_i, \ell_j) \leq b_n} \left\| K_{n,ij} \left(\frac{e_{n,i}(1)}{\pi_{n,i}} + \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \right) \left[\left(\frac{e_{n,j}(1)}{\pi_{n,j}} + \frac{e_{n,j}(0)}{1 - \pi_{n,j}} \right) [D_{n,i} D_{n,j} - \pi_{n,ij}] \right. \right. \\ & \left. \left. + 2 \frac{e_{n,j}(0)}{1 - \pi_{n,j}} (\pi_{n,i} - D_{n,i}) \right] \right\|_q \end{aligned}$$

Pulling out constants and breaking up

$$\begin{aligned} \|W_{n,i}\|_q \leq & \sum_{j: d(\ell_i, \ell_j) \leq b_n} |K_{n,ij}| \cdot \left| \frac{e_{n,i}(1)}{\pi_{n,i}} + \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \right| \cdot \left| \frac{e_{n,j}(1)}{\pi_{n,j}} + \frac{e_{n,j}(0)}{1 - \pi_{n,j}} \right| \cdot \|D_{n,i} D_{n,j} - \pi_{n,ij}\|_q \\ & + 2 \sum_{j: d(\ell_i, \ell_j) \leq b_n} |K_{n,ij}| \cdot \left| \frac{e_{n,i}(1)}{\pi_{n,i}} + \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \right| \cdot \left| \frac{e_{n,j}(0)}{1 - \pi_{n,j}} \right| \cdot \|\pi_{n,i} - D_{n,i}\|_q \end{aligned}$$

By bounded potential outcomes from Assumption 1, nondegenerate treatment probabilities from Assumption 2, $D_{n,i} \in \{0, 1\}$, and uniformly bounded $K_{n,ij}$ from Definition 9, one can find constants $c_{1,q}$ and $c_{2,q}$ that do not depend upon n (but do depend upon q) and bound the elements of the sums (see normality proof for an example of how these assumptions were used to establish such a bound). Hence

$$\|W_{n,i}\|_q \leq \sum_{j: d(\ell_i, \ell_j) \leq b_n} c_{1,q} + 2c_{2,q} \leq (c_{1,q} + 2c_{2,q}) \cdot 8(b_n/d_{\min})^2 = c_{3,q} b_n^2$$

where the bound and count on the number of units in the ball followed from Assumption 4 and Lemma 9, and where $c_{3,q}$ is yet another constant that does not depend upon n .

(Step (e)) Substitute the bound from Step (d) into the result from Step (c)

$$\begin{aligned}\text{Var} \left(\sum_{i=1}^{N_n} W_{n,i} \right) &\leq \frac{32}{d_{\min}^2} N_n b_n^2 (c_{3,2} b_n^2)^2 + 8 N_n^2 \alpha_{\infty, \infty} (b_n)^{\frac{\delta}{2+\delta}} (c_{3,2+\delta} b_n^2)^2 \\ &= c_4 N_n b_n^6 + c_5 \alpha_{\infty, \infty} (b_n)^{\frac{\delta}{2+\delta}} N_n^2 b_n^4\end{aligned}$$

where c_4 and c_5 are again constants that do not depend upon n .

(Step (f)) Chebyshev's inequality implies

$$\begin{aligned}P \left[|\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}]| \geq \nu \right] &\leq \frac{1}{\nu^2} \text{Var} (\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}]) \\ \text{By definition of } W_{n,i} &= \frac{1}{\nu^2} \text{Var} \left(\frac{1}{N_n} \sum_{i=1}^{N_n} W_{n,i} \right) \\ \text{By bound established from Step (e)} &\leq \frac{1}{\nu^2} \frac{1}{N_n^2} \left[c_4 N_n b_n^6 + c_5 \alpha_{\infty, \infty} (b_n)^{\frac{\delta}{2+\delta}} N_n^2 b_n^4 \right] \\ &= \frac{1}{\nu^2} \left[c_4 N_n^{-1} b_n^6 + c_5 \alpha_{\infty, \infty} (b_n)^{\frac{\delta}{2+\delta}} b_n^4 \right]\end{aligned}$$

Now use the rate assumptions. By [Assumption 7](#), for some $\varepsilon > 0$, $\alpha_{\infty, \infty}(m)^{\frac{\delta}{2+\delta}} \leq k_2 m^{-(4+\varepsilon)}$. Therefore,

$$\begin{aligned}P \left[|\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}]| \geq \nu \right] &\leq \frac{1}{\nu^2} \left[c_4 N_n^{-1} b_n^6 + c_5 k_2 (2b_n)^{-(4+\varepsilon)} b_n^4 \right] \\ &\leq \frac{1}{\nu^2} \left[c_4 N_n^{-1} b_n^6 + c_5 \tilde{k}_2 b_n^{-\varepsilon} \right]\end{aligned}$$

By $b_n = o(N_n^{-1/6})$, can write $b_n = k_1 N_n^{\frac{1-\eta}{6}}$ for some $\eta \in (0, 1)$ and $k_1 > 0$. Substituting in

$$\begin{aligned}P \left[|\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}]| \geq \nu \right] &\leq \frac{1}{\nu^2} \left[c_4 N_n^{-1} (k_1 N_n^{\frac{1-\eta}{6}})^6 + c_5 \tilde{k}_2 (k_1 N_n^{\frac{1-\eta}{6}})^{-\varepsilon} \right] \\ &\leq \frac{1}{\nu^2} \left[c_4 \tilde{k}_1 N_n^{-\eta} + c_5 \tilde{k}_2 N_n^{-\varepsilon \frac{1-\eta}{6}} \right]\end{aligned}$$

Given that N_n is increasing, the final expression is $o(1)$. This ensures that for any ν ,

$$P \left[|\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}]| \geq \nu \right] \rightarrow 0$$

which is another way of saying that $\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}] \xrightarrow{p} 0$.

(Step (g)) Now I show that the infeasible and feasible estimators converge to the same probability limit, i.e.

$\widehat{V}_{n,C} - \mathbb{E}[\widehat{V}_{n,C,I}] \xrightarrow{p} 0$. Recall

$$\begin{aligned}\widehat{V}_{n,C,I} &= \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} D_{n,i} D_{n,j} K_{n,ij} + \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \frac{e_{n,j}(0)}{1 - \pi_{n,j}} (1 - D_{n,i})(1 - D_{n,j}) K_{n,ij} \\ &\quad - \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} (1 - D_{n,i}) D_{n,j} K_{n,ij} - \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(0)}{1 - \pi_{n,j}} D_{n,i} (1 - D_{n,j}) K_{n,ij} \\ &= \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left(e_{n,i}(1) \frac{D_{n,i}}{\pi_{n,i}} - e_{n,i}(0) \frac{(1 - D_{n,i})}{1 - \pi_{n,i}} \right) \left(e_{n,j}(1) \frac{D_{n,j}}{\pi_{n,j}} - e_{n,j}(0) \frac{(1 - D_{n,j})}{1 - \pi_{n,j}} \right) K_{n,ij}\end{aligned}$$

Subtract the feasible from the infeasible, using that

$$\begin{aligned}e_{n,i}(1) D_{n,i} &= [Y_{n,i}(1) - \mu_{n,Y(1)}] D_{n,i} \\ \widehat{e}_{n,i}(1) D_{n,i} &= [Y_{n,i} - \widehat{\mu}_{n,Y(1)}] D_{n,i} = [Y_{n,i}(1) - \widehat{\mu}_{n,Y(1)}] D_{n,i}\end{aligned}$$

which implies the following simplified difference

$$\begin{aligned}\widehat{V}_{n,C,I} - \widehat{V}_{n,C} &= \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\left(e_{n,i}(1) \frac{D_{n,i}}{\pi_{n,i}} - e_{n,i}(0) \frac{(1 - D_{n,i})}{1 - \pi_{n,i}} \right) \left(e_{n,j}(1) \frac{D_{n,j}}{\pi_{n,j}} - e_{n,j}(0) \frac{(1 - D_{n,j})}{1 - \pi_{n,j}} \right) \right. \\ &\quad \left. - \left(\widehat{e}_{n,i}(1) \frac{D_{n,i}}{\pi_{n,i}} - \widehat{e}_{n,i}(0) \frac{(1 - D_{n,i})}{1 - \pi_{n,i}} \right) \left(\widehat{e}_{n,j}(1) \frac{D_{n,j}}{\pi_{n,j}} - \widehat{e}_{n,j}(0) \frac{(1 - D_{n,j})}{1 - \pi_{n,j}} \right) \right] K_{n,ij} \\ &= [\widehat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}] \frac{2}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left(e_{n,i}(1) \frac{D_{n,i}}{\pi_{n,i}} - e_{n,i}(0) \frac{(1 - D_{n,i})}{1 - \pi_{n,i}} \right) \left(\frac{D_{n,j}}{\pi_{n,j}} \right) K_{n,ij} \\ &\quad - [\widehat{\mu}_{n,Y(0)} - \mu_{n,Y(0)}] \frac{2}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left(e_{n,i}(1) \frac{D_{n,i}}{\pi_{n,i}} - e_{n,i}(0) \frac{(1 - D_{n,i})}{1 - \pi_{n,i}} \right) \left(\frac{(1 - D_{n,j})}{1 - \pi_{n,j}} \right) K_{n,ij}\end{aligned}$$

Therefore, I can bound

$$\begin{aligned}|\widehat{V}_{n,C,I} - \widehat{V}_{n,C}| &\leq |\widehat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}| \frac{2}{N_n} \sum_{i=1}^{N_n} \sum_{j: d(\ell_i, \ell_j) \leq b_n} c_1 - |\widehat{\mu}_{n,Y(0)} - \mu_{n,Y(0)}| \frac{2}{N_n} \sum_{i=1}^{N_n} \sum_{j: d(\ell_i, \ell_j) \leq b_n} c_2 \\ &\leq |\widehat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}| \frac{2}{N_n} N_n \frac{8(4b_n)^2}{d_{\min}^2} c_1 - |\widehat{\mu}_{n,Y(0)} - \mu_{n,Y(0)}| \frac{2}{N_n} N_n \frac{8(4b_n)^2}{d_{\min}^2} c_2\end{aligned}$$

To get the first line, I first used distance-based truncation by the kernel to rewrite the second sum. Then I used bounded potential outcomes from [Assumption 1](#), nondegenerate treatment probabilities from [Assumption 2](#),

$D_{n,i} \in \{0, 1\}$, and uniformly bounded $K_{n,ij}$ from Definition 9 to establish a bound on the elements of the sum (see normality proof for an example of how these assumptions were used to establish such a bound). To get the second line, I used rough counts on observations, which followed from Assumption 4 and Lemma 9. Further simplifying,

$$|\widehat{V}_{n,C,I} - \widehat{V}_{n,C}| \leq |\widehat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}| \cdot \widetilde{c}_1 \cdot b_n^2 - |\widehat{\mu}_{n,Y(0)} - \mu_{n,Y(0)}| \cdot \widetilde{c}_2 \cdot b_n^2$$

To get convergence, resize by the usual convergence rate $N_n^{1/2}$ and assumed rate $b_n = k_1 N_n^{\frac{1-\eta}{6}}$ for some $\eta \in (0, 1)$ and simplify:

$$|\widehat{V}_{n,C,I} - \widehat{V}_{n,C}| \leq N_n^{1/2} |\widehat{\mu}_{n,Y(1)} - \mu_{n,Y(1)}| \cdot \widetilde{c}_1 k_1^2 \cdot N_n^{-\frac{1-2\eta}{6}} - N_n^{1/2} |\widehat{\mu}_{n,Y(0)} - \mu_{n,Y(0)}| \cdot \widetilde{c}_2 k_1 \cdot N_n^{-\frac{1-2\eta}{6}}$$

Note that $N_n^{1/2} |\widehat{\mu}_{n,Y(d)} - \mu_{n,Y(d)}| = O_p(1)$. While I have not proved this directly, it can be proved in exactly the same way as asymptotic normality of $\widehat{\alpha}_{n,ATE}$. Given that this is $O_p(1) N_n^{-\frac{1-2\eta}{6}} \rightarrow 0$, convergence follows.

(Step (h)) Now that I have shown $\widehat{V}_{n,C,I} - \mathbb{E}[\widehat{V}_{n,C,I}] \xrightarrow{p} 0$, I must relate the estimator to the true randomization variance V_n , as stated in Definition 7.

First, rewrite $\mathbb{E}[\widehat{V}_{n,C,I}]$ (stated in Step (a)) as follows

$$\begin{aligned} & \mathbb{E}[\widehat{V}_{n,C,I}] \\ &= \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} [\pi_{n,ij}] K_{n,ij} + \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \frac{e_{n,j}(0)}{1 - \pi_{n,j}} [1 - \pi_{n,i} - \pi_{n,j} + \pi_{n,ij}] K_{n,ij} \\ & - \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} [\pi_{n,j} - \pi_{n,ij}] K_{n,ij} - \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(0)}{1 - \pi_{n,j}} [\pi_{n,i} - \pi_{n,ij}] K_{n,ij} \\ &= \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} [\pi_{n,ij} - \pi_{n,i}\pi_{n,j} + \pi_{n,i}\pi_{n,j}] K_{n,ij} \\ & + \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \frac{e_{n,j}(0)}{1 - \pi_{n,j}} [1 - \pi_{n,i} - \pi_{n,j} + \pi_{n,ij} - \pi_{n,i}\pi_{n,j} + \pi_{n,i}\pi_{n,j}] K_{n,ij} \\ & - \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \frac{e_{n,j}(1)}{\pi_{n,j}} [\pi_{n,j} - \pi_{n,ij} - \pi_{n,i}\pi_{n,j} + \pi_{n,i}\pi_{n,j}] K_{n,ij} \\ & - \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \frac{e_{n,i}(1)}{\pi_{n,i}} \frac{e_{n,j}(0)}{1 - \pi_{n,j}} [\pi_{n,i} - \pi_{n,ij} - \pi_{n,i}\pi_{n,j} + \pi_{n,i}\pi_{n,j}] K_{n,ij} \end{aligned}$$

Simplify

$$\begin{aligned}\mathbb{E}[\widehat{V}_{n,C,I}] &= \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{e_{n,i}(1)}{\pi_{n,i}} + \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \right] \left[\frac{e_{n,j}(1)}{\pi_{n,j}} + \frac{e_{n,j}(0)}{1 - \pi_{n,j}} \right] [\pi_{n,ij} - \pi_{n,i}\pi_{n,j}] K_{n,ij} \\ &\quad + \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} [e_{n,i}(1) - e_{n,i}(0)][e_{n,j}(1) - e_{n,j}(0)] K_{n,ij}\end{aligned}$$

Then subtract V_n from this to get

$$\begin{aligned}\mathbb{E}[\widehat{V}_{n,C,I}] - V_n &= \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} \left[\frac{e_{n,i}(1)}{\pi_{n,i}} + \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \right] \left[\frac{e_{n,j}(1)}{\pi_{n,j}} + \frac{e_{n,j}(0)}{1 - \pi_{n,j}} \right] [\pi_{n,ij} - \pi_{n,i}\pi_{n,j}] (K_{n,ij} - 1) \\ &\quad + \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} [e_{n,i}(1) - e_{n,i}(0)][e_{n,j}(1) - e_{n,j}(0)] K_{n,ij}\end{aligned}$$

The first double sum approaches zero, leaving

$$\mathbb{E}[\widehat{V}_{n,C,I}] - V_n - \frac{1}{N_n} \sum_{i=1}^{N_n} \sum_{j=1}^{N_n} [e_{n,i}(1) - e_{n,i}(0)][e_{n,j}(1) - e_{n,j}(0)] K_{n,ij} = o(1)$$

□

A.4 IV Setting

Proposition 9 Under [Assumption 8](#), can rewrite α_W using the three defined groups to break up the sums:

$$\alpha_W = \frac{\frac{1}{N_n} \sum_{i \in S_{NT}} [Y_i(1) - Y_i(0)] + \frac{1}{N_n} \sum_{i \in S_{AT}} [Y_i(1) - Y_i(0)] + \frac{1}{N_n} \sum_{i \in S_C} [Y_i(1) - Y_i(0)]}{\frac{1}{N_n} \sum_{i \in S_{NT}} [W_i(1) - W_i(0)] + \frac{1}{N_n} \sum_{i \in S_{AT}} [W_i(1) - W_i(0)] + \frac{1}{N_n} \sum_{i \in S_C} [W_i(1) - W_i(0)]}.$$

By definition of the groups, this reduces to

$$\alpha_{LATE} = \frac{\frac{1}{N_n} \sum_{i \in S_{NT}} [Y_i(1) - Y_i(0)] + \frac{1}{N_n} \sum_{i \in S_{AT}} [Y_i(1) - Y_i(0)] + \frac{1}{N_n} \sum_{i \in S_C} [Y_i(1) - Y_i(0)]}{0 + 0 + \frac{1}{N_n} \sum_{i \in S_C} [W_i(1) - W_i(0)]}$$

Under [Assumption 10](#), the numerator simplifies

$$\alpha_{LATE} = \frac{0 + 0 + \frac{1}{N_n} \sum_{i \in S_C} [Y_i(1) - Y_i(0)]}{\frac{1}{N_n} \sum_{i \in S_C} [W_i(1) - W_i(0)]}$$

Finally, note that $W_i(1) - W_i(0) = 1$ for all i in the complier group. Therefore, the denominator simply counts the number of compliers and therefore equals $|S_C|$. Cancelling $1/N_n$ in the numerator and denominator, the result follows.

Proof of Proposition 11 A Taylor expansion of the Wald ratio about the true Wald ratio implies

$$\begin{aligned} \frac{\hat{\alpha}_{n,ATE,Y}}{\hat{\alpha}_{n,ATE,W}} &= \frac{\alpha_{n,ATE,Y}}{\alpha_{n,ATE,W}} + \frac{1}{\alpha_{n,ATE,W}} (\hat{\alpha}_{n,ATE,Y} - \alpha_{n,ATE,Y}) \\ &\quad - \frac{\alpha_{n,ATE,Y}}{\alpha_{n,ATE,W}^2} (\hat{\alpha}_{n,ATE,W} - \alpha_{n,ATE,W}) + O_p \left(\left\| \begin{pmatrix} \hat{\alpha}_{n,ATE,Y} - \alpha_{n,ATE,Y} \\ \hat{\alpha}_{n,ATE,W} - \alpha_{n,ATE,W} \end{pmatrix} \right\|^2 \right) \end{aligned}$$

This can be rewritten as

$$\begin{aligned} &\hat{\alpha}_{n,IV} - \alpha_{n,IV} \\ &= \frac{\hat{\alpha}_{n,ATE,Y}}{\hat{\alpha}_{n,ATE,W}} - \frac{\alpha_{n,ATE,Y}}{\alpha_{n,ATE,W}} \\ &= \frac{(\hat{\alpha}_{n,ATE,Y} - \alpha_{n,ATE,Y}) - \alpha_{n,IV} (\hat{\alpha}_{n,ATE,W} - \alpha_{n,ATE,W})}{\alpha_{n,ATE,W}} + O_p \left(\left\| \begin{pmatrix} \hat{\alpha}_{n,ATE,Y} - \alpha_{n,ATE,Y} \\ \hat{\alpha}_{n,ATE,W} - \alpha_{n,ATE,W} \end{pmatrix} \right\|^2 \right) \end{aligned}$$

Next note that for either $X \in \{Y, W\}$, from the proof of [Proposition 4](#)

$$\hat{\alpha}_{n,ATE,X} - \alpha_{n,ATE,X} = \left\{ \frac{1}{N_n} \sum_{i=1}^{N_n} \left[\frac{X_{n,i}(0) - \mu_{n,X(0)}}{1 - \pi_{n,i}} + \frac{X_{n,i}(1) - \mu_{n,X(1)}}{\pi_{n,i}} \right] [D_{n,i} - \pi_{n,i}] \right\} (1 + o_p(1))$$

Substituting this into the above expression and rearranging,

$$\begin{aligned} &\hat{\alpha}_{n,IV} - \alpha_{n,IV} \\ &= \left\{ \frac{1}{N_n} \sum_{i=1}^{N_n} \left[\frac{e_{n,i}(1)}{\pi_{n,i}} + \frac{e_{n,i}(0)}{1 - \pi_{n,i}} \right] [D_{n,i} - \pi_{n,i}] \right\} (1 + o_p(1)) + O_p \left(\left\| \begin{pmatrix} \hat{\alpha}_{n,ATE,Y} - \alpha_{n,ATE,Y} \\ \hat{\alpha}_{n,ATE,W} - \alpha_{n,ATE,W} \end{pmatrix} \right\|^2 \right) \end{aligned}$$

where

$$e_{n,i}(d) = \frac{Y_{n,i}(d) - \alpha_{n,IV} W_{n,i}(d) - [\mu_{n,Y(d)} - \alpha_{n,IV} \mu_{n,W(d)}]}{\mu_{n,W(1)} - \mu_{n,W(0)}}$$

As in the proof of [Proposition 4](#), this is a sum of binary $D_{n,i}$ indicators (rescaled by potential outcomes and treatment probabilities), plus an extra term from the Taylor expansion. The rescaling of the $D_{n,i}$ indicators has changed, but the rescaling is still finite, and the substantive assumptions on the $D_{n,i}$ are identical. Therefore, asymptotic normality for the first portion follows analogously under the same weak dependence assumptions on treatment.

For the Taylor expansion remainder term, note that

$$\hat{\alpha}_{n,ATE,X} - \alpha_{n,ATE,X} = O_p(1/\sqrt{N_n})$$

for $X \in \{Y, W\}$, which is sufficient to guarantee

$$\sqrt{N_n} \cdot V_n^{-1/2} \cdot O_p \left(\left\| \begin{pmatrix} \hat{\alpha}_{n,ATE,Y} - \alpha_{n,ATE,Y} \\ \hat{\alpha}_{n,ATE,W} - \alpha_{n,ATE,W} \end{pmatrix} \right\|^2 \right) = O_p(1/\sqrt{N_n}) = o_p(1)$$

Appendix B

Appendix for Chapter 2

B.1 Appendix

B.1.1 Technical lemmas and proofs

Here we state and prove a technical lemma referred to in [Section 2.3.2](#), and we provide the proofs of [Propositions 13](#) and [14](#).

Lemma 11. *Assume $p, k \in \mathbb{N}$ and $p > k$. Let $\lambda \in \mathbb{R}^k$, and let $G \in \mathbb{R}^{p \times k}$ have full column rank. Let $G^\perp$ denote any $p \times (p - k)$ matrix with full column rank such that $G'G^\perp = 0_{k \times (p-k)}$. Let $\mathcal{S}_p$ denote the set of $p \times p$ symmetric positive semidefinite matrices W such that $G'WG$ is nonsingular. Then*

$$\{WG(G'WG)^{-1}\lambda : W \in \mathcal{S}_p\} = \{x : x \in \mathbb{R}^p, G'x = \lambda\} = \{G(G'G)^{-1}\lambda + G^\perp z : z \in \mathbb{R}^{p-k}\}.$$

Proof. We first show that

$$\{G(G'G)^{-1}\lambda + G^\perp z : z \in \mathbb{R}^{p-k}\} \subset \{WG(G'WG)^{-1}\lambda : W \in \mathcal{S}_p\}. \quad (\text{B.1})$$

Pick any $z \in \mathbb{R}^{p-k}$, and define

$$W(z) \equiv (G, G^\perp) \begin{pmatrix} I_k & \tilde{\lambda}z' \\ z\tilde{\lambda}' & \delta I_{p-k} \end{pmatrix} \begin{pmatrix} G' \\ G^{\perp'} \end{pmatrix}, \quad \tilde{\lambda} \equiv \frac{1}{\lambda'(G'G)^{-1}\lambda} \lambda,$$

where $\delta > 0$ is arbitrary but chosen large enough so that $W(z)$ is positive semidefinite. Then

$$W(z)G = (G, G^\perp) \begin{pmatrix} I_k & \tilde{\lambda}z' \\ z\tilde{\lambda}' & \delta I_{p-k} \end{pmatrix} \begin{pmatrix} G'G \\ 0_{(p-k) \times k} \end{pmatrix} = (G + G^\perp z\tilde{\lambda}')G'G.$$

Hence,

$$G'W(z)G = (G'G)^2,$$

implying

$$W(z)G(G'W(z)G)^{-1}\lambda = (G + G^\perp z\tilde{\lambda}')(G'G)^{-1}\lambda = G(G'G)^{-1}\lambda + G^\perp z,$$

and thus the statement (B.1) holds.

Now pick any $W \in \mathcal{S}_p$. Then $x = WG(G'WG)^{-1}\lambda$ satisfies $G'x = \lambda$. This shows that

$$\{WG(G'WG)^{-1}\lambda : W \in \mathcal{S}_p\} \subset \{x : x \in \mathbb{R}^p, G'x = \lambda\}. \quad (\text{B.2})$$

Finally, choose any $x \in \mathbb{R}^p$ satisfying $G'x = \lambda$. Since the columns of G and $G^\perp$ are (jointly) linearly independent, there exist $y \in \mathbb{R}^k$ and $z \in \mathbb{R}^{p-k}$ such that $x = Gy + G^\perp z$. Note that $\lambda = G'x = G'Gy$, so necessarily $y = (G'G)^{-1}\lambda$. We have thus shown that

$$\{x : x \in \mathbb{R}^p, G'x = \lambda\} \subset \{G(G'G)^{-1}\lambda + G^\perp z : z \in \mathbb{R}^{p-k}\}. \quad (\text{B.3})$$

The set inclusions (B.1)–(B.3) together imply the statement of the lemma. $\square$

Proof of Proposition 13. It is standard to show that (2.3) holds under Assumption 11 (Newey and McFadden, 1994). Moreover,

$$\sqrt{n} \widehat{\text{se}}(\hat{x}) \xrightarrow{p} \sum_{j=1}^p V_{jj}^{1/2} |x_j| = \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} \sqrt{x' \tilde{V} x},$$

where the equality uses Lemma 2 in Section 2.3.1, $\text{diag}(V)$ is the vector of diagonal elements of V , and we recall the notation $x = WG(G'WG)^{-1}\lambda$. Hence, letting Z denote a standard normal random variable, we

have

$$\begin{aligned}
\lim_{n \rightarrow \infty} P(r(\theta_0) \in \widehat{\text{CI}}) &= \lim_{n \rightarrow \infty} P\left(\sqrt{n}|r(\hat{\theta}) - r(\theta_0)| \leq \Phi^{-1}(1 - \alpha/2) \times \sqrt{n} \widehat{\text{se}}(\hat{x})\right) \\
&= P\left(\sqrt{x'Vx} \times |Z| \leq \Phi^{-1}(1 - \alpha/2) \times \max_{\tilde{V} \in \mathcal{S}(\text{diag}(V))} \sqrt{x'\tilde{V}x}\right) \\
&\geq P(|Z| \leq \Phi^{-1}(1 - \alpha/2)) \\
&= 1 - \alpha.
\end{aligned}$$

□

Proof of Proposition 14. Under the null hypothesis,

$$\sqrt{n}r(\hat{\theta}) \xrightarrow{d} \lambda'(G'WG)^{-1}G'WV^{1/2}Z,$$

where $V^{1/2}V^{1/2'} = V$, and $Z = (Z_1, \dots, Z_p)' \sim N(0_{p \times 1}, I_p)$. The asymptotic null distribution of the test statistic $\hat{\mathcal{T}}$ is therefore a Gaussian quadratic form:

$$n\hat{\mathcal{T}} \xrightarrow{d} Z'QZ, \quad Q \equiv V^{1/2'}WG(G'WG)^{-1}\lambda S\lambda'(G'WG)^{-1}G'WV^{1/2}.$$

Székelly and Bakirov (2003) prove that

$$P(Z'QZ \leq \text{trace}(Q) \times \tau) \geq P(Z_1^2 \leq \tau) \tag{B.4}$$

for any $p \times p$ symmetric positive semidefinite (non-null) matrix Q and any $\tau > 1.5365$. Since $(\Phi^{-1}(1 - \alpha/2))^2 > 1.5365$ for $\alpha \leq 0.215$, it follows that, under the null,

$$\begin{aligned}
P(\hat{\mathcal{T}} \leq \text{cv}_n) &\geq P\left(n\hat{\mathcal{T}} \leq \text{trace}(Q) \times (\Phi^{-1}(1 - \alpha/2))^2\right) \\
&\rightarrow P\left(Z'QZ \leq \text{trace}(Q) \times (\Phi^{-1}(1 - \alpha/2))^2\right) \\
&\geq P\left(Z_1^2 \leq (\Phi^{-1}(1 - \alpha/2))^2\right) \\
&= 1 - \alpha.
\end{aligned}$$

□

B.1.2 Simulated moments

We here argue that Proposition 13 can be extended to a setting where the model-implied moment function $h(\cdot)$ is not available analytically and we therefore resort to stochastic simulation to approximate it. The

main requirement is that the number of simulation draws M is sufficiently large relative to the empirical sample size n .

Let $\hat{h}_M(\cdot)$ denote the approximation of $h(\cdot)$ computed from M simulation draws. For example, if we are matching simple moments of the form $h(\theta) = E[\tilde{h}(\theta, \zeta)]$, where $\tilde{h}(\cdot, \cdot)$ is a deterministic function and ζ is a random vector (e.g., the shocks in the structural model), then we can set $\hat{h}_M(\theta) = \frac{1}{M} \sum_{s=1}^M \tilde{h}(\theta, \zeta_s)$, where $\zeta_1, \dots, \zeta_M$ are draws from the distribution of ζ . The minimum distance estimator $\hat{\theta}$ is defined as in (2.2), but with $\hat{h}_M(\cdot)$ in place of $h(\cdot)$. The worst-case standard errors are computed as in (2.5), except that we leave the estimator $\hat{G} = \hat{G}_M$ of $G = \frac{\partial h(\theta_0)}{\partial \theta'}$ unspecified for now (we return to this choice below).

Proposition 21. *Impose [Assumption 11](#) and $\max_j |x_j| V_{jj} > 0$. Assume further that as $n \rightarrow \infty$, we have $M \rightarrow \infty$, $n/M \rightarrow 0$, $\hat{G}_M \xrightarrow{p} G$, and $\sup_{\theta: \|\theta - \theta_0\| \leq \delta_M} \sqrt{M} \|\hat{h}_M(\theta) - h(\theta)\| = O_p(1)$ for any sequence satisfying $\delta_M \rightarrow 0$. Then the conclusions of [Proposition 13](#) apply.*

Proof. Upon inspection of the proof of [Proposition 13](#), it suffices to show that (2.3) holds. We appeal to Theorem 7.2 in [Newey and McFadden \(1994\)](#), where, in their notation, $\hat{g}_n(\theta) = \hat{h}_M(\theta) - \hat{\mu}$ and $g_0(\theta) = h(\theta) - h(\theta_0)$. The first three conditions of this theorem hold by [Assumption 11\(iii\)](#)–[\(vi\)](#). The fourth condition holds by [Assumption 11\(i\)](#) and $\sqrt{n}[\hat{h}_M(\theta_0) - h(\theta_0)] = \sqrt{n/M} \times \sqrt{M}[\hat{h}_M(\theta_0) - h(\theta_0)] = o(1) \times O_p(1)$. Finally, the fifth condition follows from

$$\begin{aligned} \sup_{\theta: \|\theta - \theta_0\| \leq \delta_M} \sqrt{n} \|\hat{g}_n(\theta) - \hat{g}_n(\theta_0) - g_0(\theta)\| &= \sup_{\theta: \|\theta - \theta_0\| \leq \delta_M} \sqrt{n} \|\hat{h}_M(\theta) - \hat{h}_M(\theta_0) - [h(\theta) - h(\theta_0)]\| \\ &\leq 2\sqrt{n/M} \times \sup_{\theta: \|\theta - \theta_0\| \leq \delta_M} \sqrt{M} \|\hat{h}_M(\theta) - h(\theta)\| \\ &= o(1) \times O_p(1). \end{aligned} \quad \square$$

Remarks.

1. The key assumption is that the number of simulation draws M is large relative to the empirical sample size n , in the sense $n/M \rightarrow 0$, so that the simulation noise in $\hat{h}_M(\cdot)$ is asymptotically negligible relative to the statistical noise in $\hat{\mu}$. In practice, we recommend that researchers choose M sufficiently large so that the simulation noise in the moment estimates $\hat{h}_M(\hat{\theta})$ is orders of magnitude smaller than the empirical moment standard errors $\hat{\sigma}_j$.
2. The uniform tightness assumption $\sup_{\theta: \|\theta - \theta_0\| \leq \delta_M} \sqrt{M} \|\hat{h}_M(\theta) - h(\theta)\| = O_p(1)$ strengthens pointwise $\sqrt{M}$ -consistency of $\hat{h}_M(\theta)$ (which usually follows straight-forwardly from a central limit theorem) to

uniform $\sqrt{M}$ -consistency for all θ in a neighborhood of θ_0 .

3. Under alternative asymptotics where M is proportional to n , the worst-case standard error formula (2.5) would need to take into account the simulation noise, and the uniform tightness assumption would need to be strengthened to stochastic equicontinuity as in Newey and McFadden (1994, Theorem 7.2).
4. Consider again the special case with $h(\theta) = E[\tilde{h}(\theta, \zeta)]$ and $\hat{h}_M(\theta) = \frac{1}{M} \sum_{s=1}^M \tilde{h}(\theta, \zeta_s)$, where we now assume that the draws $\{\zeta_s\}_{s=1}^M$ are strictly stationary across s . Then uniform tightness is implied by $\sum_{\ell=-\infty}^{\infty} \|\text{Cov}(\tilde{h}(\theta, \zeta_s), \tilde{h}(\theta, \zeta_{s-\ell}))\|$ being uniformly bounded for all θ in a neighborhood of θ_0 (this follows from Chebyshev's inequality). Notice that these conditions allow the simulated moments $\hat{h}_M(\theta)$ to be computed either by averaging across time (in a single simulated economy) or across independent simulated economies.
5. The only assumption required of the estimated Jacobian matrix $\hat{G}_M$ is consistency. Newey and McFadden (1994, Section 7.3) and Hong, Mahajan and Nekipelov (2015) discuss generic finite-difference derivative estimators based on simulation draws, and they provide assumptions ensuring consistency.

B.1.3 Details of extensions

Here we provide further details on the cases discussed in Section 2.3.3 where additional knowledge about the moment covariance matrix is available.

General constraint set. Consider a general constraint set $\tilde{\mathcal{S}}$ defined by linear equality or inequality restrictions, as well as the positive semidefiniteness constraint. In the over-identified case $p > k$, the worst-case efficient weight matrix W can be computed through two nested convex/concave optimization problems:

$$\min_{W \in \mathcal{S}_p} \max_{\tilde{V} \in \tilde{\mathcal{S}}} x(W)' \tilde{V} x(W) = \min_{z \in \mathbb{R}^{p-k}} \max_{\tilde{V} \in \tilde{\mathcal{S}}} \{G(G'G)^{-1}\lambda + G^\perp z\}' \tilde{V} \{G(G'G)^{-1}\lambda + G^\perp z\}, \quad (\text{B.5})$$

where $\mathcal{S}_p$, $x(W)$, and $G^\perp$ were defined in Section 2.3.2, and the equality follows from Lemma 11 in Appendix B.1.1. The inner maximization in (B.5) is a concave semidefinite program, as discussed in Section 2.3.3. The outer minimization is an unconstrained convex program since the objective function is a pointwise maximum of convex functions in z . The nested optimizations in (B.5) may be neither *strictly* convex nor differentiable, but for our purposes it suffices to use convex optimization algorithms that return any arbitrary local minimum z^* (which is necessarily also a global minimum). Once an optimal z^* has been

computed, a corresponding optimal weight matrix is given by the matrix $W(z^*)$ defined in the proof of [Lemma 11](#) in [Appendix B.1.1](#).

To conduct joint hypothesis tests as in [Section 2.4](#), we can simply replace the constraint set in the critical value computation [\(2.9\)](#) with $\tilde{\mathcal{S}}$.

Special case: Knowledge of the block diagonal. Suppose we know the *block* diagonal of V , while all other elements are unrestricted. That is, suppose the constraint set $\tilde{\mathcal{S}}$ is given by all symmetric positive semidefinite matrices of the form

$$V = \begin{pmatrix} V_{(1)} & ? & ? & \dots & ? \\ ? & V_{(2)} & ? & \dots & ? \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ ? & ? & \dots & ? & V_{(J)} \end{pmatrix}, \quad (\text{B.6})$$

where $V_{(j)}$ are known (or consistently estimable) square symmetric matrices (possibly of different dimensions) for $j = 1, \dots, J$. Partition the vectors x and $\hat{\mu}$ conformably as $x = (x'_{(1)}, \dots, x'_{(J)})'$ and $\hat{\mu} = (\hat{\mu}'_{(1)}, \dots, \hat{\mu}'_{(J)})'$. The worst-case asymptotic standard deviation [\(2.8\)](#), for fixed W , is then given by

$$\sqrt{\max_{\tilde{V} \in \tilde{\mathcal{S}}} x' \tilde{V} x} = \sum_{j=1}^J (x'_{(j)} V_{(j)} x_{(j)})^{1/2}. \quad (\text{B.7})$$

This follows from the same logic as in [Lemma 2](#) in [Section 2.3.1](#) once we recognize that the known block diagonal of V implies that the marginal variance of $x'_{(j)} \hat{\mu}_{(j)}$ is known for each $j = 1, \dots, J$, but the correlations among these J variables remain unrestricted. Specifically, the maximum [\(B.7\)](#) is achieved by $V = \text{Var}(\tilde{\mu})$, where the random vector $\tilde{\mu} = (\tilde{\mu}'_{(1)}, \dots, \tilde{\mu}'_{(J)})'$ has the following representation. Let $\eta = (\eta'_{(1)}, \dots, \eta'_{(J)})'$ have the covariance matrix [\(B.6\)](#), but with zeros instead of question marks. Let $\bar{\eta}$ be a scalar random variable with variance 1 that is uncorrelated with η . Then set $\tilde{\mu}_{(j)} \equiv \frac{1}{\sqrt{x'_{(j)} V_{(j)} x_{(j)}}} V_{(j)} x_{(j)} \bar{\eta} + (I - \frac{1}{x'_{(j)} V_{(j)} x_{(j)}} V_{(j)} x_{(j)} x'_{(j)}) \eta_{(j)}$, $j = 1, \dots, J$.

In the over-identified case, the worst-case efficient weight matrix can be computed by substituting the formula [\(B.7\)](#) into the nested optimization [\(B.5\)](#) (with $x = G(G'G)^{-1}\lambda + G^\perp z$).

PRICE SETTING APPLICATION: REDUCED-FORM MOMENT ESTIMATES

Moment	Estimate	Std. error $\times 1000$	Pairwise correlation		
			$E[(\Delta p)^2]$	$E[(\Delta p)^4]$	$E[\Delta p]$
Frequency	0.293	2.338	0.000	0.000	0.000
$E[(\Delta p)^2]$	0.027	0.233		0.939	0.966
$E[(\Delta p)^4]$	0.001	0.019			0.831
$E[\Delta p]$	0.145	0.754			

Table B.1: Reduced-form moment estimates and their standard errors: average weekly rate of nonzero price changes (Frequency), and moments $E[|\Delta p|^j]$, $j = 1, 2, 4$ of absolute log price changes conditional on a nonzero change. “Pairwise correlation” columns: estimated pairwise correlations across the sample moments.

B.1.4 Details of the price setting application and simulation study

Here we provide details of the data used for the empirical application in [Section 2.5.1](#), and we conduct a simulation study calibrated to this application.

Data. We use the “movement” data set for beer products (file name `wber.csv`) on the Chicago Booth website.¹ We follow [Alvarez et al. \(2016\)](#) when cleaning the data. First, we keep only data for store #122. Second, we drop any observations with prices below 20 cents or above 25 dollars (the data was collected between the years 1989 and 1994). Third, we set any absolute price changes below one cent equal to zero. Fourth, we drop the largest 1% of absolute log price changes.

[Table B.1](#) shows the $p = 4$ estimated reduced-form moments, their standard errors, and their estimated correlation matrix. The sample kurtosis (fourth moment divided by squared second moment) of log price changes equals 1.80. The zero correlation between the sample frequency of nonzero price changes (a binary outcome) and the sample moments of the price change magnitudes is mechanical. Our limited-information analysis here does not exploit this fact because we want to emulate what an applied researcher might do without thinking hard about the problem. However, the independence could be taken into account using the extensions described in [Section 2.3.3](#).

Simulation study. We apply the inference methods to data simulated from the [Alvarez and Lippi \(2014\)](#) model. The simulations treat the just-identified empirical parameter estimates (columns 1–3 in [Table 2.1](#)) as the truth, and we use a sample size of $n = 37,916$ as in the real data. The binary price change indicators are drawn i.i.d. from a binomial distribution with the model-implied success probability ([Alvarez and Lippi, 2014](#), Proposition 4). The magnitudes of the price changes are drawn from the model-implied density function

¹<https://www.chicagobooth.edu/research/kilts/datasets/dominicks>

(Alvarez and Lippi, 2014, Proposition 6).² We use 10,000 Monte Carlo repetitions. The estimation and inference procedures are the same as the ones applied to the actual data (in particular, efficient estimates are computed using the one-step approach).

Table B.2 shows that the just-identified and efficient limited-information confidence intervals have coverage probabilities very nearly equal to or exceeding the nominal level of 95% for all three parameters. Though coverage is conservative, the table shows that the average length of the confidence intervals is not more than six times that of the corresponding full-information confidence intervals. This is consistent with the empirical standard errors reported in Section 2.5.1.

Table B.2 also illustrates that the worst-case perspective is key to avoiding over-rejection in the face of limited information: Both the “efficient” t-test for the price volatility parameter and the joint Wald test of the true parameter values over-reject if we erroneously assume that all the empirical moments are independent of each other.³ In contrast, the limited-information and full-information t-tests and joint tests are correctly sized.⁴ The limited-information tests are conservative (as predicted by theory), though the joint test of parameter restrictions is only mildly conservative in this particular model.

Finally, Table B.2 shows that the limited-information efficient point estimates have slightly higher root mean squared error (RMSE) than the efficient full-information estimates. It may seem surprising at first blush that the limited-information efficient estimates can have (marginally) higher RMSE than the just-identified estimates. This is because the limited-information efficient estimates are designed to have low variance under the *worst-case* correlation structure (i.e., perfect correlation of the moments), not under the true correlation structure that is unknown to the econometrician.

B.1.5 Details of the heterogeneous agent application

We here provide further details on the application in Section 2.5.2. Table B.3 shows which impulse response moments are used to efficiently estimate the seven structural parameters, according to the moment selection procedure described in Section 2.3.2. The $p = 23$ moments are shown along the rows, while the $k = 7$ parameters are shown along the columns. A cell with an “x” indicates a non-zero efficient loading ($\hat{x}_j^*$ in the

²We simulate from this density by numerically computing the associated quantile function on a fine grid, and then passing random uniform draws through a cubic interpolation of this function.

³In the case of the volatility parameter, the independence-based efficient estimator loads positively on all four moments (i.e., $\hat{x}_j > 0$ for all j). Because three of these moments are in fact highly positively correlated with each other (see Table B.1), ignoring covariances leads to an underestimate of the standard error.

⁴We only report the joint test for the just-identified specification. This is because the joint test proposed in Section 2.4.2 requires a single choice of weight matrix, whereas the worst-case efficient point estimates of the three parameters correspond to three different choices of moments (selected as in Section 2.3.2).

MONTE CARLO SIMULATION STUDY

	Just-identified specification			Efficient specification			
	# prod.	Vol.	Menu cost	# prod.	Vol.	Menu cost	
	<u>Confidence interval coverage rate</u>						
	Full-info	94.5%	95.0%	94.7%	95.0%	95.2%	95.3%
	Independence	100.0%	95.0%	100.0%	100.0%	89.1%	100.0%
	Worst case	100.0%	99.4%	100.0%	100.0%	99.4%	100.0%
	<u>Confidence interval average length</u>						
	Full-info	0.179	0.002	0.010	0.162	0.002	0.009
	Independence	0.627	0.002	0.039	0.390	0.002	0.025
	Worst case	0.878	0.003	0.059	0.571	0.003	0.041
	<u>RMSE relative to true parameter values</u>						
	Full-info	1.53%	0.59%	0.86%	1.37%	0.58%	0.76%
	Independence	1.53%	0.59%	0.86%	1.72%	0.59%	0.99%
	Worst case	1.53%	0.59%	0.86%	1.79%	0.59%	1.03%
	<u>Rejection rate of over-identification test</u>						
	Full-info	5.01%					
	Independence	0.00%					
Worst case	0.00%						
<u>Rejection rate of joint test of true parameter values</u>							
Full-info	4.79%						
Independence	7.54%						
Worst case	2.47%						

Table B.2: Simulation results based on the empirically calibrated [Alvarez and Lippi \(2014\)](#) model. The just-identified specification uses only three moments for estimation, while the efficient specification exploits all four moments. The rows correspond to full-information inference (exploits knowledge of $\hat{V}$), inference under independence (erroneously assumes that $\hat{V}$ is diagonal), and worst-case inference (exploits only diagonal of $\hat{V}$ without assuming off-diagonal elements are zero). Estimated parameters: number of products (# prod.), volatility of desired log price (Vol.), scaled menu cost (Menu cost). The over-identification test tests the validity of the fourth non-targeted moment. The joint test of the true parameters is a Wald test (Full-info or Independence) or the test proposed in [Section 2.4.1](#) (Worst case). The nominal significance level is 5%.

notation of [Section 2.3.2](#)), while empty cells indicate zero loadings.⁵

HETEROGENEOUS AGENT APPLICATION: EFFICIENT MOMENT SELECTION

Impulse response			TR	PC	TFP			Monetary	
Var.	Shock	Horiz.			AR1	AR2	Std	AR1	AR2
TFP	TFP	0	x		x	x	x	x	
		1	x		x	x		x	
		2	x	x		x		x	
		8							
Output	TFP	0	x	x				x	
		1							
		2							
		8							
Frac	TFP	0							
		1							
		2							
		8							
Output	MP	0	x	x				x	x
		1							
		2							
		8							
Price	MP	0		x					x
		1							
		2						x	
		8							
Bond	MP	1	x	x				x	x
		2							
		8	x						x

Table B.3: Cells with an “x” indicate that the efficient estimate of the given parameter (along columns) attaches a non-zero weight to the given empirical moment (along rows). First three columns show the impulse response variable (“Var.”), shock, and quarterly horizon (“Horiz.”). Variable “Frac”: fraction of people earning less than 2/3 of GDP. Shock “MP”: monetary shock. See parameter abbreviations in [Table 2.2](#).

⁵We define a loading to be zero if $|\hat{x}_j^*| < 10^{-4}$.

Appendix C

Appendix for Chapter 3

C.1 Notation and Least Squares for the Standard Linear Model

This section defines the standard multivariate linear model, as well as results for its least squares estimation. This section serves as a building block for subsequent results involving Bayesian estimation because one can compute posterior distributions by least squares estimation (OLS or GLS) with dummy observations.

Definition 36. (Standard Normal Linear Model) Write a generic multivariate regression as

$$\underset{(T \times N)}{Y} = \underset{(T \times K)}{X} \underset{(K \times N)}{B} + \underset{(T \times N)}{\varepsilon} \quad \text{where } \varepsilon_t \sim \mathcal{N}(0, \Sigma)$$

One can also apply the vec operator to derive another useful stacking/characterization:

$$y = \text{vec}(Y) = (I_N \otimes X)\beta + \text{vec}(\varepsilon) \quad \text{vec}(\varepsilon) \sim \mathcal{N}(0, \Sigma \otimes I_T)$$

where $\beta = \text{vec}(B)$. y and β stack outcome and coefficients equation-by-equation.

The likelihood of this model is

$$L(\beta, \Sigma) \propto |\Sigma|^{-T/2} \exp \left\{ -\frac{1}{2} \text{trace} \left((Y - XB)' \Sigma^{-1} (Y - XB) \right) \right\}$$

Provided $(X'X)^{-1}$ exists so that the OLS estimator can be computed, one can also introduce $\widehat{\beta}/\widehat{B}$, complete

the square, and rewrite the above as

$$\begin{aligned}
L(\beta, \Sigma) &\propto |\Sigma|^{-(Np+M)/2} \exp \left\{ -\frac{1}{2} (\beta - \hat{\beta})' (\Sigma^{-1} \otimes X'X) (\beta - \hat{\beta}) \right\} \\
&\times |\Sigma|^{-(T-Np-M)/2} \exp \left\{ -\frac{1}{2} \text{trace} (\Sigma^{-1} (Y - X\hat{B})' (Y - X\hat{B})) \right\} \\
&\propto \mathcal{N}(\beta; \hat{\beta}, \Sigma \otimes (X'X)^{-1}) \times IW(\Sigma; (Y - X\hat{B})' (Y - X\hat{B}), T - (Np + M) - N - 1)
\end{aligned}$$

Definition 37. (Generic VARX Model) A VARX model with p lags and M additional regressors (typically including a constant) can be written

$$\underbrace{y_t}_{N \times 1} = \Phi_1 y_{t-1} + \dots + \Phi_p y_{t-p} + \underbrace{C}_{N \times M} \underbrace{z_t}_{M \times 1} + \varepsilon_t \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma)$$

This can be stacked and written as in [Definition 36](#) by taking

$$Y = \begin{pmatrix} y'_1 \\ \vdots \\ y'_T \end{pmatrix} \quad X = \begin{pmatrix} x'_1 \\ \vdots \\ x'_T \end{pmatrix} \quad x_t = \begin{pmatrix} y_{t-1} \\ \vdots \\ y_{t-p} \\ z_t \end{pmatrix} \quad \underbrace{B'}_{N \times K} = \begin{pmatrix} \Phi_1 & \dots & \Phi_p & C \end{pmatrix}$$

For prior setting, it is useful to note that the following correspondence between parameters of the VARX and parameters of the standard linear model:

$$B_{ki} = \beta_{K \cdot (i-1) + k} \quad \Phi_{\ell, ij} = B_{N \cdot (\ell-1) + j, i} = \beta_{K \cdot (i-1) + N \cdot (\ell-1) + j}$$

Proposition 22. For the standard linear model of [Definition 36](#) that assumes (conditional) homoskedasticity, OLS and GLS coincide, and their usual mean and variance formulas satisfy

$$\hat{\beta}_{OLS} = \hat{\beta}_{GLS} = \text{vec} \left((X'X)^{-1} X'Y \right) \quad V_{OLS} = V_{GLS} = \Sigma \otimes (X'X)^{-1}$$

provided $(X'X)^{-1}$ exists.

Proof. I derive expressions for GLS, and then show they coincide with OLS. First define

$$y = \text{vec}(Y) \quad \tilde{X} = I_N \otimes X \quad \tilde{\Omega} = \Sigma \otimes I_T$$

The GLS variance and estimator formulas satisfy

$$V_{GLS} = (\tilde{X}'\tilde{\Omega}^{-1}\tilde{X})^{-1}$$

$$\hat{\beta}_{GLS} = (\tilde{X}'\tilde{\Omega}^{-1}\tilde{X})^{-1}\tilde{X}'\tilde{\Omega}^{-1}y = V_{GLS}\tilde{X}'\tilde{\Omega}^{-1}y$$

Substitute in the definitions of y , $\tilde{X}$, $\tilde{\Omega}$ and simplify using Kronecker identities $(A \otimes B)^{-1} = (A^{-1} \otimes B^{-1})$ (provided A, B invertible), $(A \otimes B)(C \otimes D) = (AC \otimes BD)$, $(A \otimes B) \text{vec}(V) = \text{vec}(BV A')$ to get

$$V_{GLS} = \Sigma \otimes (X'X)^{-1} \quad \hat{\beta}_{GLS} = \text{vec}((X'X)^{-1}X'Y).$$

Notice that $\hat{\beta}_{GLS}$ is precisely $\hat{\beta}_{OLS}$. To verify the variance equivalence, write out the OLS variance formula and simplify using the same Kronecker identities:

$$V_{OLS} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\Omega}\tilde{X}(\tilde{X}'\tilde{X})^{-1} = \Sigma \otimes (X'X)^{-1}.$$

□

Definition 38. (Short and Long Regressions) Given outcomes Y and covariate blocks X_1 and X_2 , consider

1. (Long Regression): This regression exploits both covariate blocks and can be written

$$Y = X_1B_1 + X_2B_2 + \varepsilon = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} + \varepsilon$$

2. (Short Regression): This regression exploits only the first block and can be written

$$Y = X_1A + u.$$

Theorem 2. (Frisch-Waugh-Lovell Theorem) *OLS coefficients in the long regression of Definition 38 satisfy*

$$\hat{B}_i = (X_i'M_{-i}X_i)^{-1}X_i'M_{-i}Y$$

where $M_i = I - X_i(X_i'X_i)^{-1}X_i'$

If $i = 1$, then M_{-i} denotes M_2 and vice versa.

Proof. This is a standard result that can be derived from the block matrix inversion formula. $\square$

Lemma 12. (In-Sample OVB Expressions for OLS) *The OLS coefficients between the short and long regressions of Definition 38 are related via*

$$\hat{A} = \hat{B}_1 + \hat{P}\hat{B}_2 \quad \text{where} \quad \hat{P} = (X_1'X_1)^{-1}X_1'X_2$$

As a consequence, if $\hat{\varepsilon}$ and $\hat{u}$ are the in-sample OLS residuals from the long and short regression, then

$$\hat{u} = \hat{\varepsilon} + (X_2 - X_1\hat{P})\hat{B}_2 \quad (\text{C.1})$$

In words, the short regression residual equals the long regression residual plus an adjustment based on the residual information in X_2 not correlated with or captured by X_1 .

Proof. One can derive an in-sample OVB expression for the short regression coefficients via substitution using the identity $Y = X_1\hat{B}_1 + X_2\hat{B}_2 + \hat{\varepsilon}$ where $\hat{B}_i$ are the OLS coefficients and $\hat{\varepsilon}$ is the OLS residual:

$$\hat{A} = (X_1'X_1)^{-1}X_1'Y = (X_1'X_1)^{-1}X_1'(X_1\hat{B}_1 + X_2\hat{B}_2 + \hat{\varepsilon}) = \hat{B}_1 + \hat{P}\hat{B}_2 + 0 \quad (\text{C.2})$$

with the final expression using orthogonality of the OLS residuals $X_1'\hat{\varepsilon} = 0$.

Then relate the short and long regression OLS residuals by substitutions and algebra:

$$\begin{aligned} \hat{u} &= Y - X_1\hat{A} = X_1\hat{B}_1 + X_2\hat{B}_2 + \hat{\varepsilon} - X_1\hat{A} \\ &= \hat{\varepsilon} + X_1(\hat{B}_1 - \hat{A}) + X_2\hat{B}_2 \\ \text{Using C.2} \quad &= \hat{\varepsilon} + X_1(-\hat{P}\hat{B}_2) + X_2\hat{B}_2 = \hat{\varepsilon} + (X_2 - X_1\hat{P})\hat{B}_2 \end{aligned}$$

$\square$

Proposition 23. (Add-In Formula for OLS) *Consider a partitioned “full” sample split into two blocks of observations as follows:*

$$Y_f = \begin{pmatrix} Y \\ Y_d \end{pmatrix} \quad X_f = \begin{pmatrix} X \\ X_d \end{pmatrix}$$

Define the following estimators that use either the full sample or partial sample, respectively,

$$\hat{B}_f = (X'_f X_f)^{-1} X'_f Y_f \quad \hat{B}_p = (X' X)^{-1} X' Y$$

The estimators are related according to

$$\hat{B}_f = \hat{B}_p + (X' X)^{-1} X'_d (I_m + X_d (X' X)^{-1} X'_d)^{-1} [Y_d - X_d \hat{B}_p]$$

where m denotes the number of rows/observations in Y_d and X_d .

Remark. This result can be viewed as an “add-in” updating result, which is the counterpart of “leave-out” formulas familiar from cross-validation. $\hat{B}_p$ is the initial estimate based on partial information. This is updated based on how much of a “surprise”/error there is $Y_d - X_d \hat{B}_p$ for the supplemental observations when predicting Y_d using X_d and the partial sample estimator $\hat{B}_p$.

Proof. Apply the Woodbury matrix inversion formula to the following

$$\begin{aligned} (X'_f X_f)^{-1} &= (X' X + X'_d X_d)^{-1} \\ &= (X' X)^{-1} - (X' X)^{-1} X'_d (I_m + X_d (X' X)^{-1} X'_d)^{-1} X_d (X' X)^{-1} \end{aligned}$$

Use that to rewrite the expression for OLS estimator $\hat{B}_f$:

$$\begin{aligned} \hat{B}_f &= (X'_f X_f)^{-1} X'_f Y_f \\ &= [(X' X)^{-1} - (X' X)^{-1} X'_d (I_m + X_d (X' X)^{-1} X'_d)^{-1} X_d (X' X)^{-1}] [X' Y + X'_d Y_d] \\ &= [(X' X)^{-1} - (X' X)^{-1} X'_d (I_m + X_d (X' X)^{-1} X'_d)^{-1} X_d (X' X)^{-1}] [X' Y] \\ &\quad + [(X' X)^{-1} - (X' X)^{-1} X'_d (I_m + X_d (X' X)^{-1} X'_d)^{-1} X_d (X' X)^{-1}] [X'_d Y_d] \end{aligned}$$

Recognize $\widehat{B}_p$ and simplify further

$$\begin{aligned}
\widehat{B}_f &= \widehat{B}_p - (X'X)^{-1}X'_d(I_m + X_d(X'X)^{-1}X'_d)^{-1}X_d\widehat{B}_p \\
&\quad + [(X'X)^{-1}X'_dY_d - (X'X)^{-1}X'_d(I_m + X_d(X'X)^{-1}X'_d)^{-1}X_d(X'X)^{-1}X'_dY_d] \\
&= \widehat{B}_p - (X'X)^{-1}X'_d(I_m + X_d(X'X)^{-1}X'_d)^{-1}X_d\widehat{B}_p \\
&\quad + (X'X)^{-1}X'_d(I_m + X_d(X'X)^{-1}X'_d)^{-1}(I_m + X_d(X'X)^{-1}X'_d)Y_d \\
&\quad - (X'X)^{-1}X'_d(I_m + X_d(X'X)^{-1}X'_d)^{-1}X_d(X'X)^{-1}X'_dY_d \\
&= \widehat{B}_p + (X'X)^{-1}X'_d(I_m + X_d(X'X)^{-1}X'_d)^{-1} \left[-X_d\widehat{B}_p + Y_d + X_d(X'X)^{-1}X'_dY_d - X_d(X'X)^{-1}X'_dY_d \right] \\
&= \widehat{B}_p + (X'X)^{-1}X'_d(I_m + X_d(X'X)^{-1}X'_d)^{-1} \left[Y_d - X_d\widehat{B}_p \right]
\end{aligned}$$

□

C.2 Bayesian Estimation of the Linear Model

Formulas and results of this section are standard in the literature and adapted from [Kadiyala and Karlsson \(1997\)](#) and [Karlsson \(2013\)](#). Although those papers handled VARs specifically, results are stated here in general for Bayesian multivariate regression, which nests VARXs.

In this section, I characterize posteriors under increasingly relaxed priors, as well as how to compute them. In [Appendix C.2.1](#), I assume that Σ is known and derive the posteriors for coefficients. [Appendix C.2.2](#) does the same when (1) Σ is unknown but (2) the prior variance over coefficients has Kronecker structure, which is relevant for the leading equivalence result of [Section 3.3.1](#). Lastly, [Appendix C.2.3](#) dispenses with the assumption of Kronecker structure.

In each section, I describe prior implementation and estimation via OLS or GLS with dummy variables, since estimation with dummies is central to deriving the main results of this paper.

C.2.1 Coefficient Estimation in Known- or Conditional-On- Σ Case

In this section, I characterize the posterior of coefficients conditional on Σ . Although this assumption does correspond to early work on BVARs as in [Litterman \(1980\)](#), I do not intend to maintain this assumption. Rather, this case is most useful as a simpler setting in which to study the posteriors on coefficients and main ideas around dummy variable implementation of priors, ignoring uncertainty about Σ . The results in this section will also be repurposed and applied *conditional* on Σ in the posterior sampling schemes of [Appendix C.2.2](#) and [Appendix C.2.3](#).

Lemma 13. (Posterior under Generic Normal Prior, Known- Σ) *Suppose the prior $p(\beta) \sim \mathcal{N}(\beta_0, \Omega_0)$. Then the posterior is*

$$\begin{aligned}\beta|y &\sim \mathcal{N}(\bar{\beta}, \bar{\Omega}) \\ \bar{\Omega} &= [\Omega_0^{-1} + (\Sigma^{-1} \otimes X'X)]^{-1} \\ \bar{\beta} &= \bar{\Omega}[\Omega_0^{-1}\beta_0 + (\Sigma^{-1} \otimes X') \text{vec}(Y)] = \bar{\Omega}[\Omega_0^{-1}\beta_0 + (\Sigma^{-1} \otimes X'X) \text{vec}(\hat{B})]\end{aligned}$$

Remark. The second expression for $\bar{\beta}$ that exploits $\hat{B}$ is only operable if one can compute $\hat{B}$. This is not always possible if, for example, there are more regressors than observations so that $X'X$ is not invertible. However, in those cases, $\bar{\beta}$ still exists using the first expression.

Lemma 14. (Posterior under Kronecker Normal Prior, Known- Σ) *Suppose the prior is $p(\beta) \sim \mathcal{N}(\beta_0, \Sigma \otimes \Phi_0)$. Then the posterior is*

$$\begin{aligned}\beta|y &\sim \mathcal{N}(\bar{\beta}, \bar{\Omega}) & \bar{\Omega} &= \Sigma \otimes (X'X + \Phi_0^{-1})^{-1} \\ \text{unvec}(\bar{\beta}) &= \bar{B} & &= (X'X + \Phi_0^{-1})^{-1}[X'Y + \Phi_0^{-1}B_0]\end{aligned}$$

where $B_0 = \text{unvec}(\beta_0)$.

Remark. The difference from Lemma 13 is the Kronecker structure in the prior variance. The likelihood naturally has Kronecker structure, and forcing the prior to have the same induces a stronger degree of conjugacy. In particular, the posterior variance for the coefficients will also have a special Kronecker structure that is easier to compute than the posterior in Lemma 13.

Notice also that the posterior mean is precisely the ridge regression estimator, but penalizing deviations from $\Phi_0^{-1}B_0$ rather than simply zero.

The next set of results have origins back to [Theil \(1963\)](#) and [Litterman \(1980\)](#).

Proposition 24. (Implementation of Lemma 13 with Dummy Variables and GLS) *Suppose that one wishes to implement the prior of Lemma 13. Consider the stacked, vectorized system*

$$\underbrace{\begin{pmatrix} \text{vec}(Y) \\ \Omega_0^{-1/2} \beta_0 \end{pmatrix}}_{\tilde{y}} = \underbrace{\begin{pmatrix} I_N \otimes X \\ \Omega_0^{-1/2} \end{pmatrix}}_{\tilde{X}} \beta + w \quad w \sim \left(0, \underbrace{\begin{pmatrix} \Sigma \otimes I_T & 0 \\ 0 & I_{NK} \end{pmatrix}}_{\tilde{\Omega}} \right)$$

which uses the decomposition $\Omega_0 = \Omega_0^{1/2} (\Omega_0^{1/2})'$. The GLS estimator $\hat{\beta}_{GLS}$ and GLS variance V_{GLS} for this system satisfy

$$\hat{\beta}_{GLS} = (\tilde{X}' \tilde{\Omega}^{-1} \tilde{X})^{-1} \tilde{X}' \tilde{\Omega}^{-1} \tilde{y} = \bar{\beta} \quad V_{GLS} = (\tilde{X}' \tilde{\Omega}^{-1} \tilde{X})^{-1} = \bar{\Omega}$$

where $(\bar{\beta}, \bar{\Omega})$ are the mean and variance in Lemma 13.

Remark. Because this proposition works with the more general prior of Lemma 13 rather than that of Lemma 14 (which has a restrictive Kronecker structure), one must vectorize this system to estimate it via dummy variables. Said another way, the variance for the dummy observations under Lemma 13 do not share the same (homoskedastic) structure as the variance of the T genuine observations. Vectorization is necessary to appropriately estimate using dummy variables. Notice also that we must estimate via GLS (as proved below) or via OLS *after* transforming the system (see Corollary 3).

Proof. Start by writing out the GLS variance formula, making use of Kronecker identities:

$$\begin{aligned} \tilde{X}' \tilde{\Omega}^{-1} &= \begin{pmatrix} (I_N \otimes X)' (\Sigma \otimes I_T)^{-1} & \Omega_0^{-1/2} \end{pmatrix} \\ &= \begin{pmatrix} (I_N \otimes X') (\Sigma^{-1} \otimes I_T) & \Omega_0^{-1/2} \end{pmatrix} \\ &= \begin{pmatrix} (I_N \Sigma^{-1} \otimes X' I_T) & \Omega_0^{-1/2} \end{pmatrix} \\ &= \begin{pmatrix} (\Sigma^{-1} \otimes X') & \Omega_0^{-1/2} \end{pmatrix} \\ \Omega_{GLS} &= (\tilde{X}' \tilde{\Omega}^{-1} \tilde{X})^{-1} = [(\Sigma^{-1} \otimes X') (I_N \otimes X) + \Omega_0^{-1}]^{-1} \\ &= [(\Sigma^{-1} I_N \otimes X' X) + \Omega_0^{-1}]^{-1} \\ &= [(\Sigma^{-1} \otimes X' X) + \Omega_0^{-1}]^{-1} = \bar{\Omega} \end{aligned}$$

Next, write the estimator itself

$$\begin{aligned}\hat{\beta}_{GLS} &= (\tilde{X}'\tilde{\Omega}^{-1}\tilde{X})^{-1}\tilde{X}'\tilde{\Omega}^{-1}\tilde{y} = \bar{\Omega} \begin{pmatrix} (\Sigma^{-1} \otimes X') & \Omega_0^{-1/2} \end{pmatrix} \begin{pmatrix} \text{vec}(Y) \\ \Omega_0^{-1/2}\beta_0 \end{pmatrix} \\ &= \bar{\Omega}[(\Sigma^{-1} \otimes X') \text{vec}(Y) + \Omega_0^{-1}\beta_0] = \bar{\beta}\end{aligned}$$

□

Corollary 3. (Implementation of Lemma 13 with Dummy Variables and OLS) *Again suppose that one wishes to implement the prior of Lemma 13. Consider the transformed, stacked, vectorized system*

$$\begin{pmatrix} \text{vec}(Y(\Sigma^{-1/2})') \\ \Omega_0^{-1/2}\beta_0 \end{pmatrix} = \begin{pmatrix} (\Sigma^{-1/2} \otimes X) \\ \Omega_0^{-1/2} \end{pmatrix} \beta + w \quad w \sim \left(0, \begin{pmatrix} I_{NT} & 0 \\ 0 & I_{NK} \end{pmatrix}\right)$$

which uses the decompositions $\Omega_0 = \Omega_0^{1/2}(\Omega_0^{1/2})'$ and $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})'$. The OLS estimator $\hat{\beta}_{OLS}$ and OLS variance V_{OLS} for this system satisfy

$$\hat{\beta}_{OLS} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} = \bar{\beta} \quad V_{OLS} = (\tilde{X}'\tilde{X})^{-1} = \bar{\Omega}$$

where $(\bar{\beta}, \bar{\Omega})$ are the mean and variance in Lemma 13.

Proof. Simply write out

$$\begin{aligned}\begin{pmatrix} (\Sigma^{-1/2} \otimes X) \\ \Omega_0^{-1/2} \end{pmatrix}' \begin{pmatrix} (\Sigma^{-1/2} \otimes X) \\ \Omega_0^{-1/2} \end{pmatrix} &= (\Sigma^{-1/2} \otimes X)'(\Sigma^{-1/2} \otimes X) + (\Omega_0^{-1/2})'\Omega_0^{-1/2} \\ &= (\Sigma^{-1} \otimes X'X) + \Omega_0^{-1} = \bar{\Omega} \\ \begin{pmatrix} (\Sigma^{-1/2} \otimes X) \\ \Omega_0^{-1/2} \end{pmatrix}' \begin{pmatrix} \text{vec}(Y(\Sigma^{-1/2})') \\ \Omega_0^{-1/2}\beta_0 \end{pmatrix} &= (\Sigma^{-1/2} \otimes X)' \text{vec}(Y(\Sigma^{-1/2})') + (\Omega_0^{-1/2})'\Omega_0^{-1/2}\beta_0 \\ &= ((\Sigma^{-1/2})' \otimes X') \text{vec}(Y(\Sigma^{-1/2})') + (\Omega_0^{-1/2})'\Omega_0^{-1/2}\beta_0 \\ &= \text{vec}(X'Y(\Sigma^{-1/2})'\Sigma^{-1/2}) + (\Omega_0^{-1/2})'\Omega_0^{-1/2}\beta_0 \\ &= (\Sigma^{-1} \otimes X') \text{vec}(Y) + \Omega_0^{-1}\beta_0 = \bar{\Omega}^{-1}\bar{\beta}\end{aligned}$$

I obtained these expressions starting from the system in Proposition 24, and premultiplying $\tilde{y}, \tilde{X}$ by $\tilde{\Omega}^{-1/2}$.

□

Corollary 4. (Implementation of Lemma 14 with Dummy Variables and OLS) *Consider the prior in Lemma 13 for general variance-covariance matrix Ω_0 . If using the untransformed observations (Y, X) (in contrast to Corollary 3), the prior can be implemented via dummy variables and OLS if and only if Ω_0 can be written*

$$\Omega_0 = \Sigma \otimes (D'D)^{-1} \quad (\text{C.3})$$

for some $K \times K$ matrix D . Then one can define dummy variables as follows

$$\underbrace{Y_d}_{K \times N} = \underbrace{D}_{K \times K} \underbrace{B_0}_{K \times N} \quad \underbrace{X_d}_{K \times K} = D$$

and estimate the stacked, augmented system

$$\underbrace{\begin{pmatrix} Y \\ Y_d \end{pmatrix}}_{\tilde{Y}} = \underbrace{\begin{pmatrix} X \\ X_d \end{pmatrix}}_{\tilde{X}} \beta + \underbrace{\begin{pmatrix} \varepsilon \\ \varepsilon_d \end{pmatrix}}_{\tilde{\varepsilon}} \quad (\text{C.4})$$

via OLS. The resulting OLS estimator and variance expressions satisfy

$$V_{OLS} = \Sigma \otimes (X'X + D'D)^{-1}$$

$$\text{vec}^{-1}(\hat{\beta}_{OLS}) = (X'X + D'D)^{-1}(X'Y + D'DB_0)$$

As a corollary, to implement the prior of Lemma 14, one can set D based on the decomposition $\Phi_0 = (D'D)^{-1}$ and compute the posterior mean and variance via the OLS formulas.

Proof. Apply OLS to the augmented system in C.4. By Proposition 22, for any (Y_d, X_d) such that $\tilde{X}'\tilde{X}$ invertible,

$$V_{OLS} = \Sigma \otimes (\tilde{X}'\tilde{X})^{-1} = \Sigma \otimes (X'X + X_d'X_d)^{-1}$$

The above is what one gets from estimating System C.4 via OLS for some X_d choice, regardless of whether OLS can, in fact, match the target posterior variance. From Lemma 13, the correct target posterior variance

for generic Ω_0 is

$$\bar{\Omega} = [\Omega_0^{-1} + (\Sigma^{-1} \otimes X'X)]^{-1}$$

To match this posterior, equate the two expressions and isolate Ω_0 (using Kronecker identities):

$$\begin{aligned} [\Sigma \otimes (X'X + X_d'X_d)^{-1}] &= [\Omega_0^{-1} + (\Sigma^{-1} \otimes X'X)]^{-1} \\ \Omega_0 &= \Sigma \otimes (X_d'X_d)^{-1} \end{aligned}$$

Therefore, Ω_0 must satisfy Expression C.3. This takes care of the “only if.”

For the “if” part, use the stated dummy observations and verify. Again, appeal to the expressions from Proposition 22 for the OLS estimator to derive

$$\begin{aligned} V_{OLS} &= \Sigma \otimes (X'X + X_d'X_d)^{-1} \\ &= \Sigma \otimes (X'X + D'D)^{-1} \\ \text{vec}^{-1}(\hat{\beta}_{OLS}) &= (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} \\ &= (X'X + D'D)^{-1}(X'Y + D'DB_0) \end{aligned}$$

□

C.2.2 Estimation under Dependent Normal-IW Prior

The section covers the leading prior, which is used in Section 3.3.1. Unlike the previous subsection of this appendix, it allows for uncertainty about Σ .

Definition 39. (Conjugate Dependent Normal-IW Prior) This is the most tractable and widely used conjugate prior which takes form

$$\begin{aligned} p(\beta|\Sigma) &\sim \mathcal{N}(\beta_0, \Sigma \otimes \Phi_0) \\ p(\Sigma) &\sim IW(S_0, \alpha_0) \propto |\Sigma|^{-(\alpha_0+N+1)/2} \exp \left\{ -\frac{1}{2} \text{trace}(\Sigma^{-1}S_0) \right\} \end{aligned}$$

Definition 40. (Fully Diffuse Prior) One can obtain a fully-diffuse prior specification by taking the dependent Normal-IW prior, but with $\Phi_0 \rightarrow \infty$, $S_0 = 0$, $\alpha_0 = 0$.

Proposition 25. (Posterior under Conjugate Dependent Normal-IW Prior) *Under the above prior, the posterior distribution is*

$$\begin{aligned}
p(\beta|\Sigma, y) &= \mathcal{N}(\bar{\beta}, \Sigma \otimes \bar{\Phi}) \\
p(\Sigma|y) &\sim IW(\bar{S}, \bar{\alpha}) \propto |\Sigma|^{-(\bar{\alpha}+N+1)/2} \exp \left\{ -\frac{1}{2} \text{trace}(\Sigma^{-1} \bar{S}) \right\} \\
B_0 &= \text{vec}^{-1}(\beta_0) \\
\bar{\Phi} &= [\Phi_0^{-1} + X'X]^{-1} \\
\bar{B} &= \bar{\Phi}[\Phi_0^{-1}B_0 + X'Y] \\
\bar{\beta} &= \text{vec}(\bar{B}) \\
\bar{S} &= S_0 + (Y - X\hat{B})'(Y - X\hat{B}) + (B_0 - \hat{B})'(\Phi_0 + (X'X)^{-1})^{-1}(B_0 - \hat{B}) \\
&= \hat{B}'X'X\hat{B} + B_0'\Phi_0^{-1}B_0 + S_0 + (Y - X\hat{B})'(Y - X\hat{B}) - \bar{B}'\bar{\Phi}^{-1}\bar{B} \\
&= B_0'\Phi_0^{-1}B_0 + S_0 + Y'Y - \bar{B}'\bar{\Phi}^{-1}\bar{B} \\
\bar{\alpha} &= T + \alpha_0
\end{aligned}$$

Sampling can be done by first sampling $\Sigma|y$, then $\beta|\Sigma, y$.

Remark. Alternative expressions are given which do/do not make use of $\hat{B}$. This is important because $(X'X)^{-1}$ may not exist, e.g. when there are more regressors than observations.

Corollary 5. *The marginal posterior $p(B|y)$ has matrixvariate t distribution*

$$B \sim Mt_{KN}(\bar{B}, \bar{\Phi}^{-1}, \bar{S}, \bar{\alpha}) \quad \implies \quad \begin{cases} \mathbb{E}[B] &= \bar{B} \\ \text{Var}(\text{vec}(B)) &= \frac{1}{\bar{\alpha}-N-1} \bar{S} \otimes \bar{\Phi} \end{cases}$$

As a consequence, given any fixed K -dimensional vector x ,

$$x'B \sim Mt_{KN}(x'\bar{B}, [x'\bar{\Phi}^{-1}x]^{-1}, \bar{S}, \bar{\alpha} + 1) \quad \implies \quad \begin{cases} \mathbb{E}[x'B] &= x'\bar{B} \\ \text{Var}(x'B) &= \frac{x'\bar{\Phi}^{-1}x}{\bar{\alpha}-N} \bar{S} \end{cases}$$

Proof. In this proof, I make use of the definitions of distributions and a result in [Appendix C.2.4](#).

First, from Proposition [Proposition 27](#), if

$$\begin{aligned}\Sigma &\sim IW_q(Q, v) \\ X|\Sigma &\sim MN_{pq}(M, \Sigma, P)\end{aligned}$$

then

$$X \sim Mt_{pq}(M, P^{-1}, Q, \nu)$$

From Proposition [25](#), simplify identify terms:

$$Q = \bar{S} \quad v = \bar{\alpha} \quad \text{vec}(M) = \bar{\beta} \quad \Sigma = \Sigma \quad P = \bar{\Phi}$$

Next, apply Proposition [28](#). □

Remark. (On the Kronecker Form) This prior is popular because the variance in posterior $p(\beta|\Sigma, y)$ has a Kronecker structure, which is a direct consequence of assuming this particular prior $p(\beta|\Sigma)$, which also has Kronecker form in the variance. As a result, computing the posterior variance of $\beta|\Sigma, y$ is relatively efficient. This is in contrast to the independent normal-IW prior which is otherwise similar, except for the fact that it is not as restrictive in its assumptions on the prior variance, and the posterior variance is tougher to compute as a consequence.

The Kronecker structure is not without cost as it does restrict the prior distribution somewhat.¹ Namely, the prior variance-covariance matrices for coefficients in the different equations are *proportional*, and so one cannot impose, for example, more shrinkage on cross-lags than own-lags for a given equation because this would violate proportionality across equations. To see this, note that $\Sigma \otimes \Phi_0$ has blocks on the diagonal of $\Sigma_{11}\Phi_0, \dots, \Sigma_{NN}\Phi_0$, i.e. proportionality. This was a key feature of the Minnesota prior.

However, there is a Minnesota-like prior that maintains much of the spirit of the Minnesota prior, imposing more shrinkage on larger lags and more variable predictors. See [Assumption 16](#).

¹The independent normal-IW prior below relaxes this, at the expense of computational complexity.

C.2.3 Estimation under Independent Normal-IW

The dependent Normal-IW prior is convenient, but implies a particular form of dependence in the joint prior distribution of β and Σ .² This next prior recovers that flexibility by instead imposing prior independence between β and Σ (at the expense of computational complexity both in posterior sampling and posterior computations from larger matrix inversions).³

Definition 41. (Conjugate Independent Normal-IW Prior) Suppose $p(\beta, \Sigma) = p(\beta)p(\Sigma)$ with

$$p(\beta) \sim \mathcal{N}(\beta_0, \Omega_0)$$

$$p(\Sigma) \sim IW(S_0, \alpha_0) \propto |\Sigma|^{-(\alpha_0 + N + 1)/2} \exp \left\{ -\frac{1}{2} \text{trace}(\Sigma^{-1} S_0) \right\}$$

Definition 42. (Normal Diffuse Prior) This prior is the same as independent Normal-IW prior, but takes $S_0 = 0$ and $\alpha_0 = 0$. This prior is diffuse only with respect to the variance Σ , not the coefficients B .⁴

Proposition 26. (Posterior under Conjugate Independent Normal-IW Prior) *Under the above prior, one has the following conditional posterior distributions:*

$$p(\beta|\Sigma, y) \propto \mathcal{N}(\bar{\beta}, \bar{\Omega})$$

$$\bar{\Omega} = (\Omega_0^{-1} + \Sigma^{-1} \otimes X'X)^{-1}$$

$$\bar{\beta} = \bar{\Omega}[\Omega_0^{-1}\beta_0 + (\Sigma^{-1} \otimes X') \text{vec}(Y)]$$

$$= \bar{\Omega}(\Omega_0^{-1}\beta_0 + (\Sigma^{-1} \otimes X'X)\hat{\beta})$$

$$p(\Sigma|\beta, y) \propto IW(S_0 + (Y - XB)'(Y - XB), T + \alpha_0)$$

$$\propto IW(S_0 + (Y - X\hat{B})'(Y - X\hat{B}) + (B - \hat{B})'X'X(B - \hat{B}), T + \alpha_0)$$

The full posterior $p(\beta, \Sigma|y)$ is not analytical, but permits a Gibbs sampling scheme.

Note that we can also write $IW(\Psi^{-1}, d) \sim W(\Psi, d)$. This is useful because one can sample Σ^{-1} directly, rather than drawing Σ and doing a matrix inversion to draw β .

²One consequence: the prior variance for β cannot mimic the known- Σ Minnesota prior.

³There is still typically dependence in the posterior based on the data. Here, we are just referring to the prior, where independence may be justified and is commonly assumed.

⁴One could also obtain a fully diffuse prior by further taking $\Omega_0 \rightarrow \infty$, which would match the fully-diffuse prior stated above, but that started from the independent normal-IW. They both give the same limit.

C.2.4 Distributions

Definition 43. (Wishart Distribution) The $q \times q$ matrix X has Wishart distribution $X \sim W_q(V, \nu)$ if and only if it has density satisfying

$$p(X) \propto |X|^{(\nu-q-1)/2} \exp \left\{ -\frac{1}{2} \text{trace}(V^{-1}X) \right\}$$

Definition 44. (Inverse-Wishart Distribution) The $N \times N$ matrix Σ has an Inverse-Wishart distribution $IW(S, \alpha)$ if and only if it has density satisfying

$$p(\Sigma) \propto |\Sigma|^{-(\alpha+N+1)/2} \exp \left\{ -\frac{1}{2} \text{trace}(\Sigma^{-1}S) \right\}$$

Note that if $X \sim W_q(V, \nu)$, then X^{-1} has $IW_q(V^{-1}, \nu)$ distribution. Similarly, if $Y \sim IW(\Psi, \nu)$, then $Y^{-1} \sim W(\Psi^{-1}, \nu)$.

Definition 45. (Matrixvariate Normal) $B \in \mathbb{R}^{p \times q}$ has a matrixvariate normal distribution

$$B \sim MN_{pq}(M, Q, P)$$

if and only if

$$\text{vec}(B) \sim \mathcal{N}(\text{vec}(M), Q \otimes P)$$

Definition 46. (Matrixvariate t Distribution) $B \in \mathbb{R}^{p \times q}$ has a matrixvariate t distribution

$$B \sim Mt_{pq}(M, P, Q, v)$$

if and only if

$$p(B) \propto k^{-1} |Q + (B - M)'P(B - M)|^{-(v+p)/2} \implies \begin{cases} \mathbb{E}[B] &= M \\ \text{Var}(\text{vec}(B)) &= \frac{1}{\nu-q-1} Q \otimes P^{-1} \end{cases}$$

The random variable can also be defined (Dickey, 1967) as that arising from the following construction:

$$B = M + Z (V^{1/2})^{-1} \quad \text{where} \quad \begin{cases} V^{1/2}(V^{1/2})' = V \sim W(Q^{-1}, \nu) \\ Z_{\bullet, j} \stackrel{iid}{\sim} \mathcal{N}(0, P^{-1}) \end{cases}$$

Stated explicitly, Z has iid column vectors with common $\mathcal{N}(0, P^{-1})$ distribution, and V is a Wishart random variable, which is decomposed via a Cholesky decomposition into $V^{1/2}(V^{1/2})'$.

Proposition 27. (Normal-IW Result) *Suppose*

$$\begin{aligned} \Sigma &\sim IW_q(Q, v) \\ B|\Sigma &\sim MN_{pq}(M, \Sigma, P) \end{aligned}$$

Then

$$B \sim Mt_{pq}(M, P^{-1}, Q, \nu) \quad \implies \quad \begin{cases} \mathbb{E}[B] &= M \\ \text{Var}(\text{vec}(B)) &= \frac{1}{\nu-q-1} Q \otimes P \end{cases}$$

Proof. See final Appendix of Karlsson (2013). □

Proposition 28. *Suppose that*

$$B \sim Mt_{pq}(M, P, Q, \nu)$$

Then for any fixed $p \times 1$ vector x ,

$$x' B \sim Mt_{pq}(x' M, [x' P x]^{-1}, Q, \nu + 1) \quad \implies \quad \begin{cases} \mathbb{E}[B] &= x' M \\ \text{Var}(\text{vec}(B)) &= \frac{x' P x}{\nu - q} Q \end{cases}$$

Proof. By Definition 46, B can be synthesized

$$B = M + Z (V^{1/2})^{-1} \quad \text{where} \quad \begin{cases} V^{1/2}(V^{1/2})' = V \sim W(Q^{-1}, \nu) \\ Z_{\bullet, j} \stackrel{iid}{\sim} \mathcal{N}(0, P^{-1}) \end{cases}$$

Therefore,

$$x'B = x'M + x'Z (V^{1/2})^{-1}$$

Note that $x'Z$ is a row vector with iid $\mathcal{N}(0, x'P^{-1}x)$ random variables. Therefore, one could redefine

$$x'B = x'M + \tilde{Z} (V^{1/2})^{-1} \quad \text{where} \quad \begin{cases} V^{1/2}(V^{1/2})' = V \sim W(Q^{-1}, \nu) \\ Z_{\bullet, j} \stackrel{iid}{\sim} \mathcal{N}(0, x'P^{-1}x) \end{cases}$$

Simply recognize this as the distribution stated in the proposition. $\square$

C.3 Bayesian OVB Results

All adjustment procedures in this paper are motivated by Bayesian OVB results, characterized in this section. These results are most easily proved under known Σ . I discuss departures from this assumption after this initial exposition.

Given Σ , I establish OVB results depending upon whether one wishes to use a prior as in Lemma 13 or 14. These are handled in Appendix C.3.1 and Appendix C.3.2, respectively.

C.3.1 OVB Results under Prior of Lemma 13

This section develops OVB results when one wishes to implement the more general non-Kronecker conjugate prior under known Σ , stated in Lemma 13. To start, I define a reordering of covariates and coefficients in the long regression, which is convenient for analytical results by separating the two blocks of covariates distinguishing the short and long (partial and full) regressions. I maintain this reordering throughout this subsection.

Definition 47. (Reordering of Long Regression) The long regression of Definition 38 is

$$Y = \begin{pmatrix} X_1 & X_2 \end{pmatrix} B + \varepsilon \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} B_{1,1} & \cdots & B_{1,N} \\ B_{2,1} & \cdots & B_{2,N} \end{pmatrix}.$$

This is traditionally stacked/vectorized as in [Definition 36](#), which is

$$\text{vec}(Y) = \left[I_N \otimes \begin{pmatrix} X_1 & X_2 \end{pmatrix} \right] \beta + \text{vec}(\varepsilon) \quad \text{vec}(\varepsilon) \sim \mathcal{N}(0, \Sigma \otimes I_T)$$

where $\beta = \text{vec} \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} B_{1,1} \\ B_{2,1} \\ \vdots \\ B'_{1,N} \\ B'_{2,N} \end{pmatrix}.$

Notice that β stacks equation-by-equation (outcome-by-outcome), and so coefficients on the second block of regressors are mixed throughout the vector β rather than ordered at the end.

Without loss of generality, it helps this section to reorder elements in the regressor matrix and coefficient matrix. Specifically, one can define an orthogonal permutation matrix Q that reorders coefficients in $\beta = \text{vec}(B)$ so they are separated into two blocks, i.e. Q is such that

$$\begin{aligned} \text{vec}(Y) &= \left[\begin{pmatrix} I_N \otimes X_1 & I_N \otimes X_2 \end{pmatrix} \right] \tilde{\beta} + \text{vec}(\varepsilon) \\ \begin{pmatrix} I_N \otimes X_1 & I_N \otimes X_2 \end{pmatrix} &= \left[I_N \otimes \begin{pmatrix} X_1 & X_2 \end{pmatrix} \right] Q' \\ \tilde{\beta} = Q\beta &= \begin{pmatrix} \text{vec}(B_1) \\ \text{vec}(B_2) \end{pmatrix} = \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = \begin{pmatrix} B_{1,1} \\ \vdots \\ B_{1,N} \\ B_{2,1} \\ \vdots \\ B_{2,N} \end{pmatrix}. \end{aligned}$$

Finally, if one has a prior over coefficients under the usual stacking $p(\beta_0, \Omega_0)$, there is a corresponding prior over the reordered coefficient vector

$$\tilde{\beta} = Q\beta \sim \mathcal{N}(Q\beta_0, Q\Omega_0Q')$$

Therefore, one can immediately translate the prior despite reordering.

Remark. In the above, I used $\tilde{\beta}$ to denote the reordered coefficient vector. In results to come, I assume

coefficients have been reordered, but omit the additional notation to avoid clutter.

Definition 48. (Priors for Short and Long Regressions) Consider the reordered long regression of [Definition 47](#) and assume one wishes to implement a prior for β in this regression according to [Lemma 13](#) for some choice of prior mean and variance (β_0, Ω_0) . Partition

$$\Omega_0 = \begin{pmatrix} \Omega_{0,11} & \Omega'_{0,21} \\ \Omega_{0,21} & \Omega_{0,22} \end{pmatrix} \quad \beta_0 = \begin{pmatrix} \beta_{0,1} \\ \beta_{0,2} \end{pmatrix}$$

Notice that this separates the prior into blocks: one for each set of coefficients. For the short regression, one uses only $\Omega_{0,11}$ and $\beta_{0,1}$.

Finally, because Ω_0 is symmetric positive semidefinite, it admits a Cholesky decomposition

$$\Omega_0 = LL' \quad \text{where} \quad L = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \implies L^{-1} = \begin{pmatrix} L_{11}^{-1} & 0 \\ -L_{22}^{-1}L_{21}L_{11}^{-1} & L_{22}^{-1} \end{pmatrix}$$

where the expression for L^{-1} followed from the block matrix inversion formula.

Given L , one also has the following relationship between the blocks of Ω_0 :

$$\Omega_0 = \begin{pmatrix} \Omega_{0,11} & \Omega'_{0,21} \\ \Omega_{0,21} & \Omega_{0,22} \end{pmatrix} = LL' = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} L'_{11} & L'_{21} \\ 0 & L'_{22} \end{pmatrix} = \begin{pmatrix} L_{11}L'_{11} & L_{11}L'_{21} \\ L_{21}L'_{11} & L_{21}L'_{21} + L_{22}L'_{22} \end{pmatrix}$$

Note also that

$$\Omega_0^{-1} = (LL')^{-1} = (L')^{-1}L^{-1} = (L^{-1})'L^{-1} \implies \Omega_0^{-1/2} = L^{-1}$$

Proposition 29. (Bayesian Estimation of Long Regression) *Suppose that one wishes to estimate the long regression using the prior of [Lemma 13](#) for some Ω_0 and β_0 , ordered and partitioned as in [Definition 48](#).*

First, one can derive the posterior for β by OLS using the original sample augmented with dummy observations as follows:

$$y_* = \begin{pmatrix} \text{vec}(Y(\Sigma^{-1/2})') \\ L_{11}^{-1}\beta_{0,1} \\ -L_{22}^{-1}L_{21}L_{11}^{-1}\beta_{0,1} + L_{22}^{-1}\beta_{0,2} \end{pmatrix} \quad X_* = \begin{pmatrix} X_{*,1} & X_{*,2} \end{pmatrix} = \begin{pmatrix} (\Sigma^{-1/2} \otimes X_1) & (\Sigma^{-1/2} \otimes X_2) \\ L_{11}^{-1} & 0 \\ -L_{22}^{-1}L_{21}L_{11}^{-1} & L_{22}^{-1} \end{pmatrix}$$

Second, the ability to estimate the posterior using OLS (with dummy variables) implies that the posterior mean satisfies

$$\bar{\beta} = \begin{pmatrix} \bar{\beta}_1 \\ \bar{\beta}_2 \end{pmatrix} = (X'_* X_*)^{-1} X'_* y_* = \begin{pmatrix} (X'_{*,1} M_{*,2} X_{*,1})^{-1} X'_{*,1} M_{*,2} y_* \\ (X'_{*,2} M_{*,1} X_{*,2})^{-1} X'_{*,2} M_{*,1} y_* \end{pmatrix} \quad (\text{C.5})$$

where

$$M_{*,i} = I - X_{*,i} (X'_{*,i} X_{*,i})^{-1} X'_{*,i}$$

Proof. One can implement the prior of Lemma 13 via dummy variables, transformed observations, and OLS as in Corollary 3. To get there, start by defining dummy variables

$$y_d = \Omega_0^{-1/2} \beta_0, \quad X_d = \Omega_0^{-1/2}.$$

But notice that, by the Cholesky decomposition for Ω_0 in Definition 48, $\Omega_0^{-1/2} = L^{-1}$, hence

$$\begin{aligned} \Omega_0^{-1/2} &= \begin{pmatrix} L_{11}^{-1} & 0 \\ -L_{22}^{-1} L_{21} L_{11}^{-1} & L_{22}^{-1} \end{pmatrix} \implies X_d = \begin{pmatrix} L_{11}^{-1} & 0 \\ -L_{22}^{-1} L_{21} L_{11}^{-1} & L_{22}^{-1} \end{pmatrix}, \\ y_d &= \begin{pmatrix} L_{11}^{-1} \beta_{0,1} \\ -L_{22}^{-1} L_{21} L_{11}^{-1} \beta_{0,1} + L_{22}^{-1} \beta_{0,2} \end{pmatrix}. \end{aligned}$$

Therefore, the vectorized long regression with dummy variables can be written

$$\begin{pmatrix} \text{vec}(Y) \\ y_d \end{pmatrix} = \begin{pmatrix} (I_N \otimes X_1) & (I_N \otimes X_2) \\ & X_d \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + w \quad w \sim \left(0, \underbrace{\begin{pmatrix} \Sigma \otimes I_T & 0 \\ 0 & I_{N \cdot K} \end{pmatrix}}_{\Omega_*} \right).$$

This needs to be estimated by GLS to derive the posterior mean. But following the same logic as Corollary 3, one can standardize all observations (both true and dummy observations) to get a homoskedastic system by transforming the system. We cannot immediately apply Corollary 3 because of the reordering of covariates,

but we can follow the same proof steps, premultiplying the system by

$$\Omega_*^{-1/2} = \begin{pmatrix} \Sigma^{-1/2} \otimes I_T & 0 \\ 0 & I_{N \cdot K} \end{pmatrix}$$

which made use of decomposing Σ (for example via Cholesky) as $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})'$. After premultiplying, that gives y_* and X_* as stated in the proposition. $\square$

Proposition 30. (Bayesian Estimation of Short Regression) *One can derive the posterior mean under the prior of Lemma 13 with $\Omega_0 = \Omega_{0,11}$ and $\beta_0 = \beta_{0,1}$ defined in Definition 48 by using OLS on the original sample augmented with dummy observations as follows:*

$$y_* = \begin{pmatrix} \text{vec}(Y(\Sigma^{-1/2})') \\ L_{11}^{-1}\beta_{0,1} \end{pmatrix} \quad X_* = \begin{pmatrix} (\Sigma^{-1/2} \otimes X_1) \\ L_{11}^{-1} \end{pmatrix}$$

Proof. From Proposition 3, one can implement the prior of Lemma 13 via dummy variables and OLS. To do so in this case, one must append the dummy variables

$$y_d = \Omega_{0,11}^{-1/2} \beta_{0,1} \quad X_d = \Omega_{0,11}^{-1/2}$$

To show that L_{11}^{-1} can serve as $\Omega_{0,11}^{-1/2}$, recall the original Cholesky decomposition of Ω_0 in Definition 48, which implies

$$\Omega_{0,11} = L_{11}L_{11}' \implies \Omega_{0,11}^{-1} = (L_{11}^{-1})'L_{11}^{-1} \implies \Omega_{0,11}^{-1/2} = L_{11}^{-1}$$

Substituting into the expressions for y_d and X_d , the definitions of y_* and X_* follow. Then invoke Corollary 3. $\square$

Theorem 3. (In-Sample Bayesian OVB) *The posterior mean of the short regression implied by Proposition 30 is*

$$\bar{a} = (\Omega_{0,11}^{-1} + (\Sigma^{-1} \otimes X_1'X_1))^{-1}(\Omega_{0,11}^{-1}\beta_{0,1} + (\Sigma^{-1} \otimes X_1') \text{vec}(Y))$$

The posterior mean of the long regression implied by Proposition 29 is

$$\bar{\beta} = \begin{pmatrix} \bar{\beta}_1 \\ \bar{\beta}_2 \end{pmatrix} = (\Omega_0^{-1} + (\Sigma^{-1} \otimes X'X))^{-1} (\Omega_0^{-1}\beta_0 + (\Sigma^{-1} \otimes X') \text{vec}(Y)).$$

The posterior means are related by the following Bayesian OVB formula:

$$\bar{a} = \underbrace{\bar{\beta}_1}_{\text{Coefficients on } X_1} + \underbrace{P}_{\text{Projection Coefficient}} \underbrace{\bar{\beta}_2}_{\text{Coefficients on } X_2} + \underbrace{R}_{\text{Cross-Coefficient Correlation}} \quad (\text{C.6})$$

where

$$\begin{aligned} \Omega_0 &= \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix}' \\ X_d &= -L_{22}^{-1}L_{21}L_{11}^{-1} \\ y_d &= X_d\beta_{0,1} + L_{22}^{-1}\beta_{0,2} \\ X_* &= \begin{pmatrix} X_{*,1} & X_{*,2} \end{pmatrix} = \begin{pmatrix} (\Sigma^{-1/2} \otimes X_1) & (\Sigma^{-1/2} \otimes X_2) \\ L_{11}^{-1} & 0 \\ X_d & L_{22}^{-1} \end{pmatrix} \\ V &= ((\Sigma^{-1} \otimes X_1'X_1) + L_{11}^{-1'}L_{11}^{-1})^{-1} = ((\Sigma^{-1} \otimes X_1'X_1) + \Omega_{0,11}^{-1})^{-1} \\ R &= -VX_d'(I_{K_2} + X_dVX_d')^{-1}[y_d - X_d\bar{a}] \end{aligned}$$

In addition,

$$\underbrace{y_* - X_{*,1}(\bar{a} - R)}_{u_*} = \underbrace{y_* - X_{*,1}\bar{\beta}_1 - X_{*,2}\bar{\beta}_2}_{\varepsilon_*} + (X_{*,2} - X_{*,1}P)\bar{\beta}_2 \quad (\text{C.7})$$

where

$$\begin{aligned} y_* &= \begin{pmatrix} \text{vec}(Y(\Sigma^{-1/2})') \\ L_{11}^{-1}\beta_{0,1} \\ X_d\beta_{0,1} + L_{22}^{-1}\beta_{0,2} \end{pmatrix} \\ P &= (X_{*,1}'X_{*,1})^{-1}X_{*,1}'X_{*,2} \end{aligned}$$

Remark. Notice that if $L_{21} = 0$, which corresponds to the case where Ω_0 is block-diagonal, then $R = 0$. Further simplifications also obtain if Ω_0 has Kronecker structure as in Lemma 14. See Theorem 4.

Proof. By Proposition 29 and the definitions, OLS regression of y_* on $(X_{*,1}, X_{*,2})$ as defined in this theorem equals the posterior mean $\bar{\beta}' = (\bar{\beta}'_1, \bar{\beta}'_2)'$.

Next, define $\bar{a}_L$ as the OLS coefficients obtained from regressing y_* on $X_{*,1}$ (omitting $X_{*,2}$). Because this is just OLS, one can invoke the OVB formula of Lemma 12 to derive

$$\bar{a}_L = \bar{\beta}_1 + P \bar{\beta}_2 \quad \text{where} \quad P = (X'_{*,1} X_{*,1})^{-1} X'_{*,1} X_{*,2} \quad (\text{C.8})$$

Next, comparing this last regression of y_* on $X_{*,1}$ (delivering $\bar{a}_L$) against the regression with dummy variables in the short regression of Proposition 30 (delivering $\bar{a}$), notice that the only difference is the addition of extra dummy variables to handle any cross-coefficient correlation in the prior. As a result, $\bar{a}_L$ equals $\bar{a}$ plus an adjustment to account for leaving out the extra variables. The adjustment follows by applying Proposition 23, delivering

$$\bar{a}_L = \bar{a} + V X'_d (I_{K_2} + X_d V X'_d)^{-1} [y_d - X_d \bar{a}] \quad (\text{C.9})$$

where y_d and X_d are the extra (dummy) observations. Explicitly,

$$X_d = -L_{22}^{-1} L_{21} L_{11}^{-1} \quad y_d = X_d \beta_{0,1} + L_{22}^{-1} \beta_{0,2}$$

Rearrange and define

$$\bar{a} = \bar{a}_L + \underbrace{(-V X'_d (I_{K_2} + X_d V X'_d)^{-1} [y_d - X_d \bar{a}])}_R \quad (\text{C.10})$$

Into this last expression, replace $\bar{a}_L$ with Equation (C.8) to get Equation (C.6).

Finally, to relate residuals between models, again apply Lemma 12,

$$\underbrace{y_* - X_{*,1} \bar{a}_L}_{u_*} = \underbrace{y_* - X_{*,1} \bar{\beta}_1 - X_{*,2} \bar{\beta}_2}_{\varepsilon_*} + (X_{*,2} - X_{*,1} P) \bar{\beta}_2 \quad (\text{C.11})$$

This is essentially Equation (C.7), with a minor rewrite to replace $\bar{a}_L$ with $\bar{a} - R$, according to Equation (C.10). $\square$

C.3.2 OVB Results under Prior of Lemma 14

This section develops OVB results under the conjugate Kronecker prior with known Σ , stated in Lemma 14. The results and structure of this section mirror Appendix C.3.1. But because this is used for the main equivalence result of Section 3.3.1, I still state results explicitly.

Definition 49. (Priors for Short and Long Regressions) Consider the long regression of Definition 38 and assume that one wishes to implement a prior for this regression according to Lemma 14 for some choice of (invertible) Φ_0 and B_0 . Partition this prior as

$$\Phi_0 = \begin{pmatrix} \Phi_{0,11} & \Phi'_{0,21} \\ \Phi_{0,21} & \Phi_{0,22} \end{pmatrix} \quad B_0 = \begin{pmatrix} B_{0,1} \\ B_{0,2} \end{pmatrix}$$

Because Φ_0 is symmetric positive semidefinite, it admits a lower Cholesky decomposition

$$\Phi_0 = LL' \quad \text{where} \quad L = \begin{pmatrix} L_{11} & 0 \\ L_{12} & L_{22} \end{pmatrix}$$

Note that all the relations in Definition 48 hold, but with Φ_0 replacing Ω_0 .

Proposition 31. (Bayesian Estimation of Long Regression) Suppose one wishes to estimate the long regression using the prior of Lemma 14 for some Φ_0, B_0 , partitioned as in Definition 49.

First, one can derive the posterior mean for B by OLS using the original sample augmented with dummy observations as follows:

$$Y_* = \begin{pmatrix} Y \\ L_{11}^{-1}B_{0,1} \\ -L_{22}^{-1}L_{12}L_{11}^{-1}B_{0,1} + L_{22}^{-1}B_{0,2} \end{pmatrix} \quad X_* = \begin{pmatrix} X_{*,1} & X_{*,2} \end{pmatrix} = \begin{pmatrix} X_1 & X_2 \\ L_{11}^{-1} & 0 \\ -L_{22}^{-1}L_{12}L_{11}^{-1} & L_{22}^{-1} \end{pmatrix}$$

Second, estimating the posterior using OLS (with dummy variables) implies posterior mean

$$\bar{B} = \begin{pmatrix} \bar{B}_1 \\ \bar{B}_2 \end{pmatrix} = (X'_*X_*)^{-1}X'_*Y_* = \begin{pmatrix} (X'_{*,1}M_{*,2}X_{*,1})^{-1}X'_{*,1}M_{*,2}Y_* \\ (X'_{*,2}M_{*,1}X_{*,2})^{-1}X'_{*,2}M_{*,1}Y_* \end{pmatrix} \quad (\text{C.12})$$

where

$$M_{*,i} = I - X_{*,i}(X'_{*,i}X_{*,i})^{-1}X'_{*,i}$$

Proof. From Corollary 4, one can implement the prior using dummy observations built from D satisfying $\Phi_0^{-1} = D'D$. To find a suitable D , use the Cholesky decomposition of Φ_0 in Definition 49 and the fact that

$$\Phi_0^{-1} = (LL')^{-1} = (L')^{-1}L^{-1} = (L^{-1})'L^{-1}.$$

Therefore, take $D = L^{-1}$. Then simply use the block matrix inversion formula on L :

$$D = L^{-1} = \begin{pmatrix} L_{11}^{-1} & 0 \\ -L_{22}^{-1}L_{12}L_{11}^{-1} & L_{22}^{-1} \end{pmatrix}.$$

Use this D to define dummy observations as stated in Corollary 4. By that corollary and Theorem 2, Expression C.12 follows as an expression for the posterior mean. $\square$

Proposition 32. (Bayesian Estimation of Short Regression) *Suppose one wishes to estimate the short regression defined in Definition 38 under the prior of Lemma 14 with $\Phi_0 = \Phi_{0,11}$ and $B_0 = B_{0,1}$. One can derive the posterior mean by OLS using the original sample augmented with dummy observations as follows*

$$\begin{pmatrix} Y \\ L_{11}^{-1}B_{0,1} \end{pmatrix} \quad \begin{pmatrix} X_1 \\ L_{11}^{-1} \end{pmatrix}$$

Proof. From Corollary 4, one can implement the prior using dummy observations built from D satisfying $\Phi_{0,11}^{-1} = D'D$. To find a suitable D , use the Cholesky decomposition of Φ_0 in Definition 49 and the relationship implied between the blocks of Φ_0 and the blocks of L :

$$\begin{pmatrix} \Phi_{0,11} & \Phi'_{0,21} \\ \Phi_{0,21} & \Phi_{0,22} \end{pmatrix} = \begin{pmatrix} L_{11}L'_{11} & L_{11}L'_{12} \\ L_{12}L'_{11} & L_{12}L'_{12} + L_{22}L_{22} \end{pmatrix}$$

Notice that $\Phi_{0,11} = L_{11}L'_{11}$. Therefore, $\Phi_{0,11}^{-1} = (L_{11}^{-1})'L_{11}^{-1}$. Therefore, one can take $D = L_{11}^{-1}$ and construct dummy observations according to Corollary 4. $\square$

Theorem 4. (In-Sample OVB Expressions for Bayesian/Ridge Regression) *The posterior mean estimate*

implied by Proposition 32 is

$$\bar{A} = (X_1' X_1 + \Phi_{0,11}^{-1})^{-1} (X_1' Y + \Phi_{0,11}^{-1} B_{0,1})$$

The posterior mean estimate implied by Proposition 31 is

$$\bar{B} = \begin{pmatrix} \bar{B}_1 \\ \bar{B}_2 \end{pmatrix} = (X' X + \Phi_0^{-1})^{-1} (X' Y + \Phi_0^{-1} B_0).$$

The coefficients are related by the following Bayesian OVB formula:

$$\bar{A} = \underbrace{\bar{B}_1}_{\text{Coefficients on } X_1} + \underbrace{P}_{\text{Projection Coefficient}} \underbrace{\bar{B}_2}_{\text{Coefficients on } X_2} + \underbrace{R}_{\text{Cross-Coefficient Correlation}} \quad (\text{C.13})$$

where

$$\begin{aligned} \Phi_0 &= \begin{pmatrix} L_{11} & 0 \\ L_{12} & L_{22} \end{pmatrix} \begin{pmatrix} L_{11} & 0 \\ L_{12} & L_{22} \end{pmatrix}' \\ X_d &= -L_{22}^{-1} L_{12} L_{11}^{-1} \\ Y_d &= X_d B_{0,1} + L_{22}^{-1} B_{0,2} \\ X_* &= \begin{pmatrix} X_{*,1} & X_{*,2} \end{pmatrix} = \begin{pmatrix} X_1 & X_2 \\ L_{11}^{-1} & 0 \\ X_d & L_{22}^{-1} \end{pmatrix} \\ V &= (X_1' X_1 + L_{11}^{-1'} L_{11}^{-1})^{-1} = (X_1' X_1 + \Phi_{0,11}^{-1})^{-1} \\ R &= -V X_d' (I_{K_2} + X_d V X_d')^{-1} [Y_d - X_d \bar{A}] \end{aligned}$$

In addition,

$$\underbrace{Y_* - X_{*,1}(\bar{A} - R)}_{u_*} = \underbrace{Y_* - X_{*,1}\bar{B}_1 - X_{*,2}\bar{B}_2}_{\varepsilon_*} + (X_{*,2} - X_{*,1}P)\bar{B}_2 \quad (\text{C.14})$$

where

$$Y_* = \begin{pmatrix} Y \\ L_{11}^{-1} B_{0,1} \\ X_d B_{0,1} + L_{22}^{-1} B_{0,2} \end{pmatrix}$$

$$P = (X'_{*,1} X_{*,1})^{-1} X'_{*,1} X_{*,2}$$

Remark. Notice $R = 0$ if $L_{12} = 0$, which corresponds to the case where Φ_0 is block-diagonal.

Proof. By Proposition 31 and the definitions, OLS regression of Y_* on $(X_{*,1}, X_{*,2})$ delivers $\bar{B}$. This is true regardless of whether $\bar{B}$ was, in fact, computed with dummy observations or analytically. The equivalence still holds.

Next, define $\bar{A}_L$ as the OLS coefficients obtained from regressing Y_* on $X_{*,1}$ (omitting $X_{*,2}$). One can then invoke the OVB formula of Lemma 12

$$\bar{A}_L = \bar{B}_1 + P \bar{B}_2 \quad \text{where} \quad P = (X'_{*,1} X_{*,1})^{-1} X'_{*,1} X_{*,2} \quad (\text{C.15})$$

Next, comparing the regression of Y_* on $X_{*,1}$ (delivering $\bar{A}_L$) against the regression with dummy variables in Proposition 32 (delivering $\bar{A}$), notice that the only difference is the addition of extra dummy variables to handle any cross-coefficient correlation in the prior. As a result, $\bar{A}_L$ equals $\bar{A}$ plus an adjusted to account for leaving out the extra variables. Specifically, applying Proposition 23,

$$\bar{A}_L = \bar{A} + V X'_d (I_{n_2 p + m_2} + X_d V X'_d)^{-1} [Y_d - X_d \bar{A}] \quad (\text{C.16})$$

where Y_d and X_d are the extra (dummy) observations. Explicitly,

$$X_d = -L_{22}^{-1} L_{12} L_{11}^{-1} \quad Y_d = X_d B_{0,1} + L_{22}^{-1} B_{0,2}$$

Rearrange and define

$$\bar{A} = \bar{A}_L - \underbrace{V X'_d (I_{K_2} + X_d V X'_d)^{-1} [Y_d - X_d \bar{A}]}_R$$

Into this last expression, replace $\bar{A}_L$ with Equation (C.15) to get Equation (C.13).

Finally, again applying Lemma 12,

$$\underbrace{Y_* - X_{*,1}\bar{A}_L}_{u_*} = \underbrace{Y_* - X_{*,1}\bar{B}_1 - X_{*,2}\bar{B}_2}_{\varepsilon_*} + (X_{*,2} - X_{*,1}P)\bar{B}_2 \quad (\text{C.17})$$

This is essentially Equation (C.14), with a minor rewrite to replace $\bar{A}_L$ with $\bar{A} - R$, according to Equation (C.16). $\square$

C.3.3 Results for Adjustment Procedures

This section contains further details on proofs for the adjustment procedures of Section 3.3.

Proof. (Proposition 17) This proof about the result of Definition 26 ultimately relies on Theorem 4. This theorem applies in the setting of Proposition 17 with Σ uncertainty, even though I proved Theorem 4 for the posterior mean of coefficients under the prior of Lemma 14 that *ignored* Σ uncertainty. This is because the conditional-on- Σ posterior mean of coefficients is independent of Σ and coincides with that under Lemma 14; therefore, the marginals coincide as well. See Corollary 5 for further reference.

I now proceed through each step of Definition 26 to then show that the tilting done in Step 4 recenters the posterior predictive (forecast) distribution from the core VARX to instead match the full-information VARX.

1. Nothing to prove, but note that all constructions in this step of Definition 26 are motivated by Theorem 4.
2. The regression in this step simply computes P in Theorem 4. Therefore, given the definition of $\hat{v}_*$ in Step 3,

$$\begin{aligned} \hat{v}_* &= X_{*,2} - X_{*,1}P = X_{*,2} - X_{*,1}(X'_{*,1}X_{*,1})^{-1}X'_{*,1}X_{*,2} \\ &= (I - X_{*,1}(X'_{*,1}X_{*,1})^{-1}X'_{*,1})X_{*,2} = M_{*,1}X_{*,2} \end{aligned}$$

where the last line defines the annihilator matrix with respect to $X_{*,1}$.

3. (Key Step): For the OLS regression of $\hat{u}_*$ on $\hat{v}_* = M_{*,1}X_{*,2}$, first simply write out

$$\begin{aligned} H &= (\hat{v}_*' \hat{v}_*)^{-1} \hat{v}_*' \hat{u}_* \\ &= (X_{*,2}' M_{*,1}' M_{*,1} X_{*,2})^{-1} X_{*,2}' M_{*,1}' \hat{u}_* \\ &= (X_{*,2}' M_{*,1} X_{*,2})^{-1} X_{*,2}' M_{*,1}' \hat{u}_* \end{aligned}$$

where the last line followed because $M_{*,1}$ is idempotent. Next, use the definition of $\hat{u}_*$ built in Step 1 of [Definition 26](#):

$$\begin{aligned} H &= (X_{*,2}' M_{*,1} X_{*,2})^{-1} X_{*,2}' M_{*,1}' (Y_* - X_{*,1}(\bar{A} - R)) \\ &= (X_{*,2}' M_{*,1} X_{*,2})^{-1} X_{*,2}' M_{*,1}' Y_* - 0 \end{aligned}$$

which followed since $M_{*,1}X_{*,1} = 0$. Notice that this is just $\bar{B}_2$ by [Proposition 31](#).

4. (Adjust the Forecast): Step 2 pinned down P , and Step 3 showed that $H = \bar{B}_2$. Therefore, the constraint $\bar{g}$ in Step 4 can be written

$$\bar{g} = x_{T+1,1}(\bar{A} - R) + (x_{T+1,2} - x_{T+1,1}P)\bar{B}_2$$

Use the OVB expression for $\bar{A}$ from [Theorem 4](#):

$$\begin{aligned} \bar{g} &= x_{T+1,1}(\bar{B}_1 + P\bar{B}_2 + R - R) + [x_{T+1,2} - x_{T+1,1}P]\bar{B}_2 \\ &= x_{T+1,1}\bar{B}_1 + x_{T+1,2}\bar{B}_2 \end{aligned}$$

Notice that this is identical to marginal posterior predictive (forecast) distribution for y_{T+1} that one obtains from the full-information VARX, from [Proposition 15](#).

By construction, the mean $\mathbb{E}_*[y_{T+1}]$ of the tilted distribution will equal $\bar{g}$, the mean of the forecast distribution for the full-information VARX. Therefore, the forecast that emerges from the procedure of [Definition 26](#) is mean-equivalent to re-estimation of the full model. $\square$

Proof. (Further Details for [Proposition 19](#)) I claimed in the body that, for a given draw of $\Sigma_A^{(d)}$, the procedure

of [Definition 28](#) establishes a map between.

$$\mathbb{E}_A[a|\Sigma_A = \Sigma, Y] = \mathbb{E}_A[\text{vec}(A)|\Sigma_A = \Sigma, Y] \quad \rightarrow \quad \mathbb{E}_B[b|\Sigma_B = \Sigma, Y] = \mathbb{E}_B[\text{vec}(B)|\Sigma_B = \Sigma, Y]$$

I now justify that claim.

In the above proof of [Proposition 17](#), I showed that [Theorem 4](#) justifies the four steps of [Definition 26](#). Similarly, conditional on a value of Σ , [Theorem 3](#) justifies the four steps of [Definition 28](#). The proof is done exactly the same way, walking through each step. Focusing on Step 4 in particular, note that

$$\begin{aligned} g^{(d)} &= (I_N \otimes x_{T+1,1})(\bar{a}^{(d)} - R) + h((I_N \otimes x_{T+1,2}) - (I_N \otimes x_{T+1,1})P) \\ &= x_{T+1,1}(A^{(d)} - \text{vec}^{-1}(R)) + ((I_N \otimes x_{T+1,2}) - (I_N \otimes x_{T+1,1})P)H \\ &= x_{T+1,1}A^{(d)} - x_{T+1,1}(\text{vec}^{-1}(R) - \text{vec}^{-1}(PH)) + x_{T+1,2}H \end{aligned}$$

Conditional on a value of Σ , the four steps in [Definition 28](#). One can repeat them with minor adjustments and draw the analogous conclusions, conditional on a $\Sigma_A^{(d)}$ draw. In the body, I discussed the implications when marginalizing over Σ_A draws. $\square$

C.4 Additional Prior Setting Info for BVARs

Assumption 24. (Original Minnesota Prior) The original Minnesota prior of [Litterman \(1980\)](#) makes three key assumptions:

1. Known/fixed $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$ with diagonal form where the elements of σ_i^2 are given. In practice, these are not known but are set to the residual variance from separate univariate AR models estimated on each variable, with the same order p as the VAR.

This puts inference in the setting of [Lemma 13](#), where the econometrician specifies $p(\beta) \sim \mathcal{N}(\beta_0, \Omega_0)$.

2. $\beta_0 = \mathbb{E}[\beta]$ set as in [Assumption 14](#).
3. $\Omega_0 = \text{Var}(\beta)$ set as in [Assumption 15](#).

Remark. (Inspecting the Minnesota-Like Prior) [Assumption 16](#) maintains much of the spirit of the Minnesota prior of [Litterman \(1980\)](#) in that it does impose more shrinkage on larger lags and more variable predictors. However, because of its use Kronecker structure in the variance of coefficients, one cannot maintain all features of the priors in [Assumption 24](#) or [Assumption 15](#). I now explore the prior in a bit more

detail. Throughout, recall that β represents a stacking of—in order—all coefficients for equation/endogenous variable 1, then all coefficients for equation/endogenous variable 2, etc.

Given a value for Σ , write out the conditional prior variance of β :

$$\begin{aligned} \text{Var}(\beta|\Sigma) &= \Sigma \otimes \Phi_0 \\ &= \begin{pmatrix} \Sigma_{11}\Phi_0 & \cdots & \Sigma_{1n}\Phi_0 \\ \vdots & \ddots & \vdots \\ \Sigma_{N1}\Phi_0 & \cdots & \Sigma_{NN}\Phi_0 \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{11}\Phi_{0,11} & \cdots & 0 & \cdots & \Sigma_{1n}\Phi_{0,11} & \cdots & 0 \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_{11}\Phi_{0,KK} & \cdots & 0 & \cdots & \Sigma_{1n}\Phi_{0,KK} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \Sigma_{N1}\Phi_{0,11} & \cdots & 0 & \cdots & \Sigma_{NN}\Phi_{0,11} & \cdots & 0 \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_{N1}\Phi_{0,KK} & \cdots & 0 & \cdots & \Sigma_{NN}\Phi_{0,KK} \end{pmatrix} \end{aligned}$$

Next, note that the *marginal* prior distribution for β will have variance-covariance matrix of $(\alpha_0 - N - 1)^{-1}(S_0 \otimes \Phi_0)$.⁵ Therefore, under [Assumption 16](#)'s choice of, Φ_0, S_0, α_0 ,

$$\frac{1}{\alpha_0 - N - 1} S_0 \otimes \Phi_0 = \begin{pmatrix} \left(\frac{\lambda_1}{\ell^{\lambda_3}}\right)^2 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\sigma_1^2}{\sigma_N^2} \left(\frac{\lambda_1}{\ell^{\lambda_3}}\right)^2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \cdots & \frac{\sigma_N^2}{\sigma_1^2} \left(\frac{\lambda_1}{\ell^{\lambda_3}}\right)^2 & \cdots & 0 \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & \left(\frac{\lambda_1}{\ell^{\lambda_3}}\right)^2 \end{pmatrix}$$

This last display mimics Ω_0 of [Assumption 15](#), except for the fact that it does not include cross-variable shrinkage as a direct result of the Kronecker structure.

⁵See [Karlsson \(2013\)](#).

C.5 Empirical Application

C.5.1 Data

Sources

- [1] Real-time data from Philadelphia Fed [website](#).
- [2] The FRED [website](#) maintained by the Federal Reserve Bank of St. Louis, for many of the non-realtime series.
- [3] The census [website](#) for supplemental housing price deflator data

Core Time Series The core time series are

- Growth of real GDP, computed as 400 times the log difference of real GDP to get annualized percentage point change. This series is built from data in the file `routputMvQd.xlsx`, downloaded from [1].
- Inflation in GDP deflator, computed as 400 times the log difference of the GDP price index. This series is built from data in the file `pMvQd.xlsx`, downloaded from [1].
- Unemployment rate in units of percentage points, with the quarterly variable formed as within-quarter averages of the constituent monthly unemployment rates. This series is built from monthly data in the file `unempl.xlsx`, downloaded from [1].
- Three-month T-Bill rate in percentage points, annualized term. The quarterly rates are computed as within-quarter averages of the constituent monthly average rate, which is the TB3MS series downloaded from [2].

Auxiliary Data for GDP Growth

- Prior quarter growth of real GDP, which is simply the lag of the same series used in the core model.
- Monthly employment growth, computed as 1200 times the log difference of employment growth. This series is built from the data in the file `emp.xlsx`, downloaded from [1]
- The level of the ISM production index for manufacturing for the first month of the current quarter. This data source is not publicly available but can be downloaded from subscription services providing access to economic data series.

Auxiliary Data for GDP Inflation

- Monthly inflation in the CPI ex food and energy, computed as 1200 times the log difference of the index. This index is the series CPILFESL from [2].
- Monthly inflation in the CPI for food, computed as 1200 times the log difference of the index. This index is the series CPIUFDSL from [2].
- Monthly inflation in the CPI for energy, computed as 1200 times the log difference of the index. This index is the series CPIENGSL from [2].
- Krüger et al. (2017) call for inflation in the “PPI for capital goods.” For this, I use monthly inflation the series WPSFD41312 from [2], labeled as the “Producer Price Index by Commodity: Final Demand: Private Capital Equipment.”
- Monthly inflation in the price deflator index for new single-family houses under construction. This index is labeled “Monthly - Single-Family Houses Under Construction,” contained in the file `price_uc_cust.xls`, downloaded from [3].

Auxiliary Data for Unemployment

- Monthly unemployment rate, in units of percentage points. This is the same series used in the core model, but not averaged to produce a quarterly series.
- Monthly employment growth, computed as 1200 times the log difference of employment growth. This series is built from the data in the file `emp.xlsx`, downloaded from [1], and is the same as that used for output.
- Monthly growth in new claims for unemployment insurance, computed as 1200 times the log difference of the series ICSA from [2].

Auxiliary Data for T-Bill rate

- Monthly average of the 3-month T-Bill rate in percentage points, for an annualized term, which is available directly as the TB3MS series downloaded from [2].
- The mid-month 3-month T-Bill rate in percentage points, for an annualized term. The daily 3-month T-Bill rate is available as DTB3 from [2]. From the daily data, the mid-month value can be extracted. I choose either the 15th or, if that is not a business day, the first available next business day.

- The mid-month 6-month T-Bill rate in percentage points, for an annualized term. The daily 6-month T-Bill rate is available as DTB6 from [2]. This series uses same convention to extract as the mid-month 3-month T-Bill rate.

C.5.2 Additional Figures for Nowcasts

Nowcast of GDP Growth For predicting GDP growth according to Model 3.10, the scatterplots below plot GDP growth against the different predictors used in the nowcast regression. The red dot indicates the final observation 2019Q1 (actual for the predictor, and predicted for GDP growth), while the black dots are data prior to that point.

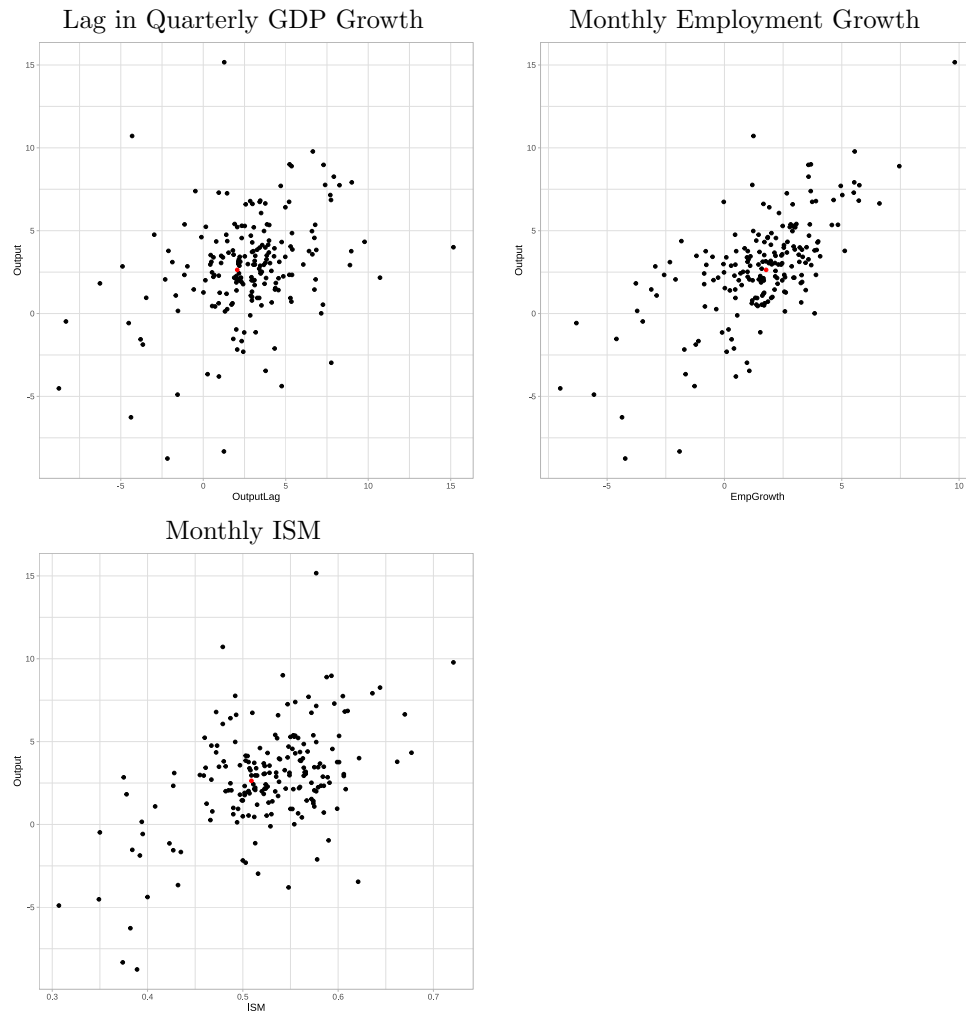


Figure C.1: Scatterplots showing GDP growth (y-axis) plotted against regressors used in constructing the nowcast of GDP growth (x-axis).

Nowcast of GDP Inflation For predicting GDP growth according to Model 3.10, the scatterplots below plot GDP growth against the different predictors used in the nowcast regression. The red dot indicates the final observation 2019Q1 (actual for the predictor, and predicted for GDP inflation), while the black dots are data prior to that point.

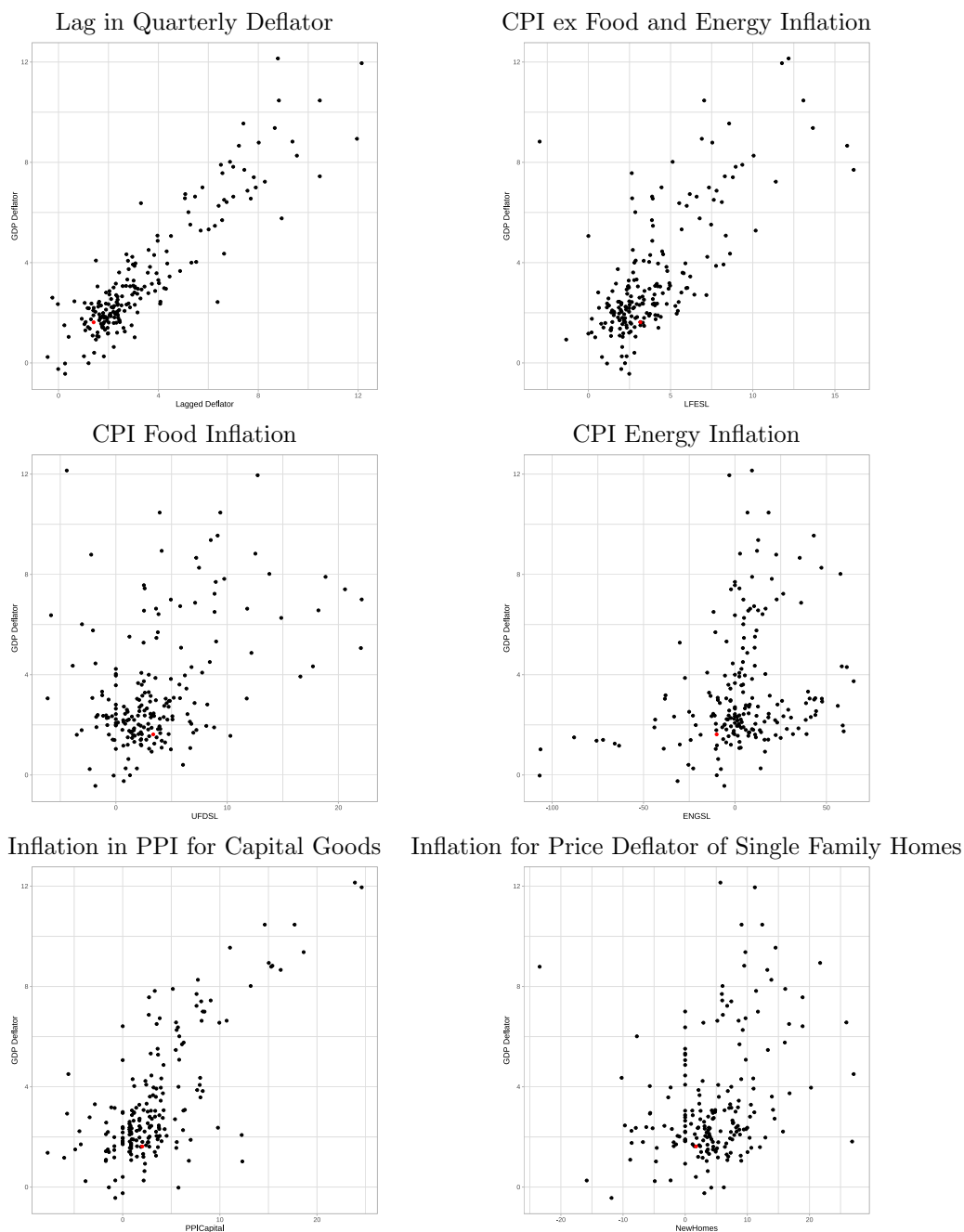


Figure C.2: Scatterplots showing GDP growth (y-axis) plotted against regressors used in constructing the nowcast of GDP growth (x-axis).

Nowcast of Unemployment The figures below show the monthly series and example monthly forecasts used to construct the nowcast of quarterly unemployment.

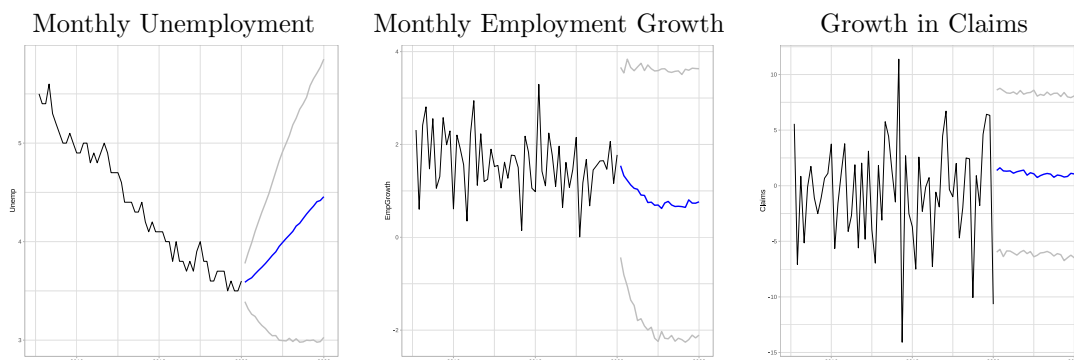


Figure C.3: Example forecasts from the monthly model used to form the nowcast of quarterly unemployment.

Nowcast of TBill The figures below show the monthly series and example monthly forecasts used to construct the nowcast of the average TBill over the quarter.

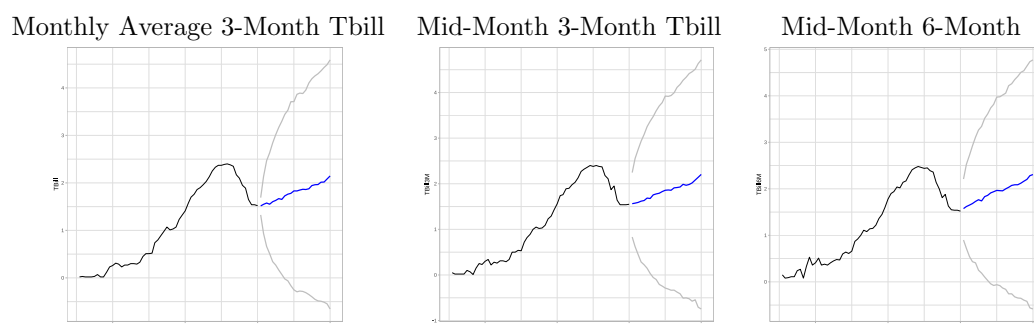


Figure C.4: Example forecasts from the monthly model used to form the nowcast of quarterly unemployment.

C.5.3 Additional Empirical Results

Effect of Sample

Out of concern for time-variation in the dynamics of macro aggregates over the sample, [Krüger et al. \(2017\)](#) and [Carriero et al. \(2015b\)](#) consider nowcasting under stochastic volatility. Therefore, as a robustness check, I consider the effect of changing the sample starting period for estimation of the nowcast models. Rather than start in 1970:Q1 (default), I begin the sample in 1985:Q1, which means discarding data before the start of the Great Moderation.

The table below shows performance when tilting based on nowcasts (only) constructed with the two alternative samples. The choice of a shorter sample does not lead to uniformly greater forecast. Performance is mixed and any improvements are below 15 basis points.

	Sample	GDP Growth	Inflation	Unemployment	T-Bill
h = 1	Default	1.8	1.02	0.13	0.060
	Short	1.9	0.92	0.13	0.088
h = 2	Default	2.4	1.10	0.33	0.41
	Short	2.4	1.03	0.33	0.46
h = 3	Default	2.5	1.21	0.53	0.75
	Short	2.5	1.15	0.53	0.78
h = 4	Default	2.5	1.33	0.74	1.05
	Short	2.5	1.29	0.75	1.08

Effect of Lag

As mentioned in the body in [Section 3.4.2](#), [Krüger et al. \(2017\)](#)'s description seems to omit the lag of GDP inflation in that variable's nowcasting model. However, I include it to have symmetry with the GDP growth case and because it does make a difference empirically, as the table below shows. The evaluation period reported below runs from 2010:Q1-2018:Q4. Notice also that even though the inclusion of the lag only directly affects forecasts in the first period, the improved accuracy also has large implications for accuracy of the T-Bill rate multiple periods ahead. This indicates that the core BVAR does capture an important dynamic relationship between inflation and the interest rate, and accurate forecast and tilting in the first period translates into better performance even in outer horizons.

	Lag	GDP Growth	Inflation	Unemployment	T-Bill
h = 1	No	1.5	1.9	0.13	0.02
	Yes	1.4	0.9	0.13	0.02
h = 2	No	1.6	1.5	0.27	0.17
	Yes	1.7	0.9	0.25	0.09
h = 3	No	1.7	1.4	0.37	0.35
	Yes	1.7	0.9	0.34	0.21
h = 4	No	1.8	1.5	0.47	0.51
	Yes	1.7	0.8	0.42	0.35

Comparison of Subsamples

This is similar to Table 3.2 in all aspects, except it breaks up the sample period into subsamples.

	Model	GDP Growth	Inflation	Unemployment	Tbill
1993:Q1-2000:Q4					
h = 1	(1) BVAR	2.2	0.87	0.17	0.33
	(2) Nowcasts	1.6	0.87	0.13	0.08
	(3) Procedure	1.8	0.70	0.10	0.15
h = 2	(1) BVAR	2.3	1.20	0.25	0.63
	(2) Nowcasts	2.2	1.20	0.23	0.45
	(3) Procedure	2.2	1.10	0.21	0.48
h = 3	(1) BVAR	2.3	1.40	0.33	0.87
	(2) Nowcasts	2.3	1.40	0.28	0.75
	(3) Procedure	2.4	1.30	0.27	0.76
h = 4	(1) BVAR	2.2	1.80	0.44	1.20
	(2) Nowcasts	2.3	1.70	0.37	1.10
	(3) Procedure	2.3	1.70	0.38	1.10
2001:Q1-2008:Q4					
h = 1	(1) BVAR	2.7	1.20	0.29	0.49
	(2) Nowcasts	2.2	1.20	0.13	0.08
	(3) Procedure	1.9	0.94	0.14	0.13
h = 2	(1) BVAR	3.4	1.20	0.60	0.91
	(2) Nowcasts	3.3	1.20	0.43	0.51
	(3) Procedure	3.3	1.10	0.44	0.56
h = 3	(1) BVAR	3.4	1.30	1.00	1.30
	(2) Nowcasts	3.4	1.20	0.81	0.94
	(3) Procedure	3.4	1.30	0.81	0.97
h = 4	(1) BVAR	3.3	1.40	1.40	1.600
	(2) Nowcasts	3.5	1.30	1.20	1.300
	(3) Procedure	3.4	1.30	1.20	1.300
2009:Q1-2018:Q4					
h = 1	(1) BVAR	2.3	0.98	0.24	0.24
	(2) Nowcasts	1.4	0.99	0.13	0.03
	(3) Procedure	1.5	0.87	0.12	0.06
h = 2	(1) BVAR	2.4	0.96	0.36	0.40
	(2) Nowcasts	2.1	0.94	0.29	0.18
	(3) Procedure	2.2	0.92	0.27	0.22
h = 3	(1) BVAR	2.6	0.99	0.53	0.56
	(2) Nowcasts	2.4	0.93	0.40	0.36
	(3) Procedure	2.4	0.93	0.39	0.39
h = 4	(1) BVAR	2.8	0.89	0.66	0.75
	(2) Nowcasts	2.6	0.87	0.52	0.55
	(3) Procedure	2.6	0.95	0.52	0.59

Table C.1: Comparison of RMSE. The BVAR (1) represents unconditional forecasts under the first prior, as described in Section 3.4.2. The accuracy for (2) and (3) is the accuracy of the adjustment procedures.

This is similar to Table 3.4 in all aspects, except it breaks up the sample period into subsamples.

	Model	GDP Growth	Inflation	Unemployment	Tbill
1993:Q1-2000:Q4					
h = 1	(1) BVAR	1.9	0.82	0.17	0.37
	(2) Nowcasts	1.6	0.89	0.14	0.07
	(3) Procedure	1.7	0.82	0.11	0.21
h = 2	(1) BVAR	2.0	1.1	0.22	0.73
	(2) Nowcasts	1.9	1.2	0.25	0.44
	(3) Procedure	2.0	1.1	0.21	0.55
h = 3	(1) BVAR	2.1	1.3	0.28	1.00
	(2) Nowcasts	2.1	1.3	0.32	0.82
	(3) Procedure	2.1	1.3	0.27	0.93
h = 4	(1) BVAR	2.0	1.6	0.38	1.4
	(2) Nowcasts	2.1	1.6	0.42	1.2
	(3) Procedure	2.1	1.7	0.38	1.3
2001:Q1-2008:Q4					
h = 1	(1) BVAR	2.7	1.2	0.32	0.44
	(2) Nowcasts	2.1	1.2	0.13	0.07
	(3) Procedure	2.3	1.1	0.14	0.10
h = 2	(1) BVAR	3.3	1.1	0.58	0.82
	(2) Nowcasts	3.2	1.1	0.43	0.47
	(3) Procedure	3.3	1.2	0.45	0.48
h = 3	(1) BVAR	3.3	1.2	0.97	1.20
	(2) Nowcasts	3.5	1.2	0.79	0.88
	(3) Procedure	3.4	1.3	0.81	0.89
h = 4	(1) BVAR	3.2	1.4	1.3	1.5
	(2) Nowcasts	3.5	1.3	1.1	1.2
	(3) Procedure	3.3	1.4	1.2	1.2
2009:Q1-2018:Q4					
h = 1	(1) BVAR	2.0	0.92	0.21	0.10
	(2) Nowcasts	1.5	0.93	0.13	0.03
	(3) Procedure	1.8	0.90	0.12	0.03
h = 2	(1) BVAR	2.0	0.92	0.30	0.20
	(2) Nowcasts	1.9	0.94	0.27	0.13
	(3) Procedure	2.0	0.92	0.27	0.15
h = 3	(1) BVAR	2.0	0.90	0.40	0.31
	(2) Nowcasts	2.0	0.95	0.37	0.28
	(3) Procedure	2.0	0.91	0.36	0.30
h = 4	(1) BVAR	2.0	0.84	0.47	0.43
	(2) Nowcasts	2.1	0.85	0.45	0.41
	(3) Procedure	2.1	0.88	0.44	0.41

Table C.2: Comparison of RMSE. The BVAR (1) represents unconditional forecasts under the second prior, as described in [Section 3.4.2](#). The accuracy for (2) and (3) is the accuracy of the adjustment procedures.