Essays in Theoretical and Applied Econometrics

by

Aibo Gong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2022

Doctoral Committee:

Professor Matias D. Cattaneo, Co-Chair
Assistant Professor Andreas Hagemann, Co-Chair
Professor Xuming He
Associate Professor Shaowei Ke
Professor Melvin Stephens Jr

Aibo Gong

aibogong@umich.edu

ORCID iD:0000-0002-6739-5954

To my parents, 巩振山 and 董代弟

ACKNOWLEDGEMENTS

First and foremost, I am deeply indebted to my advisor, mentor, and friend, Matias Cattaneo. Without his continued support and encouragement, this dissertation will not be possible, and it is also doubtful whether I can succeed in getting my Ph.D. degree. His endless patience, energy, and valuable suggestion have been the most encouraging and supportive things to confront all these difficulties throughout my Ph.D. life. As a colleague, working with him is a fantastic experience. I benefit a lot not only from his knowledge, but also from his philosophy about research. All these help shape my mind as a researcher and guide me the direction in my future academic journey. It is one of the most fortunate, if not the only, things to have such an advisor who always puts things that benefit students most at the first priority.

I am also fortunate enough to have the opportunity to communicate with all these fantastic faculties at University of Michigan. I am lucky to have a great dissertation committee. I am grateful for my dissertation commmittee co-chair, Andreas Hagemann, who not only provides valuable advice and feedback about my job market paper, but also helps me handle all these daily issues during the difficult time in pandemic. I would like to thank professor Shaowei Ke, my close friend and collaborator, who also contributes a lot to my job market paper as well as my first published work. It is a really fantastic knowing and working together with him and it is so lucky of me to become a friend with him during my first year here. As a friend, he is always there and willing to help whenever needed. As a coauthor, we start working together since 2017 and I benefit a lot from his energy and wisdom. I would like to say thank you to professor Xuming He and professor Mel Stephens, who read my job market paper and sit through my practice job talks multiple times and provide

valuable feedback and advice during my job market.

I am super grateful to Xinwei Ma, Yingjie Feng, Gonzalo Vazques-Bare, Kenichi Nagasawa for their continuous help during their study at University of Michigan and even after the graduation. From discussing possible research ideas to getting advice about the job market paper and all the issues during the job market, it is really great to have some people in higher cohorts who provide selfless help. These weekly phone calls from Xinwei during the pandemic that discuss daily life to research advice really make the time staying home not that boring and unbearable. I would also say thank you to Sebastian Calonico, Max Farrell and Michael Jansson for their helpful comments. I am grateful to Yuehao Bai and Florian Gunsilius. Knowing someone who is close and works on similar things during the pandemic is really helpful to get rid of the blue caused by staying at home. Having someone who understands your struggle during the process of producing the job market paper and encourages you to complete it is really a great thing while preparing for all these chapters in my dissertation.

I would also like to express my gratitude to many others. Thanks to Xuehang Fu for making me feel being cared and relieving all these unreasonable anxieties during the most difficult time. Without you, this dissertation will not exist. Thanks to my roommate Jinsun Liu. I cannot imagine how my life in Ann Arbor, especially the two-year life that was mostly spent at home during the pandemic, will be without such a great roommate. Thanks to Ming Fang and Lunyang Huang for these remote self study in zoom during the pandemic and talks that support each other during the difficult times in Ph.D. study. Thanks to Zheng Gong and Yian Yin for all these mock interviews and useful suggestions during the job market. Thanks to Tianbai Wang for all these phone calls when I am in such blue. Thanks to Yanliang Li and Dota2, my friend and the game that makes my Ph.D. life not completely boring. Thanks to my friends, Lu Gan, Rebecca Gao, Chen Li, Zhihan Liu, Yining Lu, Jiyuan Song, Lingchen Sun, Shuqiao Sun, Wenbo Sun, Xuan Teng, Zhongda Wang, Yijun Wu, Huayu Xu, Yinghan Xu, Hang Yu, Junming Zhang, Liang Zhu, Xinping Zuo, who helped me in many ways and made these years really great and enjoyable experience.

Last, I would like to say thank you to my parents. They may never know what my research is exact about and what econometrics mean. However, their unreserved supports and continuous encouragement are the most powerful things that support me during the Ph.D. study. It has been quite a long time before we last met each other because of the pandemic, and it has been a longer time that I spent abroad and did not stay with them. It is not an easy decision and I really appreciate their sacrifice and support all these years.

Without the help and kindness of all these people around, mentioned or maybe carelessly missed, the dissertation could not come out and I could not finish my Ph.D. study and have the chance to continue my academic journey. Wish you all the best. Hope this is not the ending point of my research life and hope this is not only a hope.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Despite their wide use in empirical applications, traditional econometric tools may perform poorly in applied work, as the difficulties faced by researchers in applied work are often overlooked through reasonings that depend on restrictive conditions. This dissertation consists of three connected chapters on essential issues in conducting robust estimation and causal inference for key economic parameters under different setups.

The first chapter discusses identification and estimation issues on the treatment effect with anticipation, a generalization of widely used stringent assumptions. Potential outcomes frameworks with assumptions motivated by economic models are provided and bounds for treatment effects are achived. Correpsonding estimation and inference procedures are provided, as well as generalizations to incorporate complicated situations to achieve improvement over current practice.

The second chapter provides estimation and inference procedures robust to high-dimensional covariates in an important class of broadly applied cluster models. Robustness is achieved through either generalization of heteroskedasticity consistent estimators or the leave-one-out procedure.

The third chapter studies a strategic trading model between a market maker who behaves like an "econometrician" and uses econometric tools to price and a well-informed inside trader. We focus on the application of econometric tools in estimating unknown parameters in a model that is robust to information ambiguity. Unique linear equilibrium exhibits the underreaction phenomenon. We also show the equivalence between a robust linear strategy and a specific two-way learning procedure regardless of the statistical models chosen by the market maker.

CHAPTER I

Introduction

As a field that is important in both theory and application, econometrics provide people with tools to better understand models and data. However, traditional econometric tools may perform poorly in applications, as conditions required to achieve theoretical results may not be consistent with the situations faced by applied researchers. It is important to provide more robust frameworks to better deal with empirical problems and models. This dissertation concerns conducting robust estimation and causal inference for key economic parameters under various setups, either through the improvement of existing econometric tools or the introduction of econometric tools to new models.

Chapter II analyzes the treatment effect with the existence of anticipatory behavior. The concept of anticipation is not unfamiliar to researchers in economics and social sciences. It occurs when forward-looking units change their behaviors in reaction to the possibility of a new policy, and a treatment thus has an impact before its implementation. Instead of assuming 'no anticipation' like other papers, I propose analysis of the treatment effect robust to anticipation. In this chapter, I employ a potential outcomes framework and propose partial identification, estimation and inference strategies for the treatment effect robust to the presence of anticipation. I start with a classical difference-in-differences model with two time periods and provide partially identified sets with easy-to-implement estimation and inference strategies for causal parameters. Modifications to incorporate more empir-

ically relevant situations and empirical applications are also included to show how this method can be used in applied work.

Following a similar logic, chapter III presents novel cluster robust inference techniques for linear regression models with many controls. Researchers often include many covariates in their linear regression models to control for confounders in empirical research in economics, statistics and social sciences. It is also common practice in empirical work to use cluster-robust standard errors. Chapter III, joint with Matias Cattaneo, Michael Jansson and Whitney K. Newey develops inference methods that are robust to the presence of many covariates and to clustering. We find that when the number of included covariates grows at a rate related to both cluster and sample sizes, the Liang-Zeger and HC-k cluster robust standard errors are invalid in general. We propose two cluster robust standard error formulas that are robust to the inclusion of many covariates. One follows the spirit of the "Hamard" estimator studied by Cattaneo, Jansson, and Newey (2018) and the other follows the spirit of the leave-out estimator. These standard errors are also valid when the regressors and error terms cluster at different levels and when the cluster size is not ignorable. In particular, we highlight important inference problems related to clustering used in current practice.

Chapter IV concerns an applied theory model where the robustness is achieved by introducing econometric tools into this framework. In this chapter, joint with Shaowei Ke, Rui Shen and Yawen Qiu, we study strategic trading with a market maker who does not know the joint distribution of public information and an assets value, and hence cannot interpret information properly. Following a public event, liquidity traders and a probabilistically informed trader who knows the distribution trade. The market maker, who behaves like an "econometrician" with some basic statistical knowledge, adopts a robust pricing strategy that has the best worst-case payoff guarantee to estimate the parameter unknown to him. We show that such a strategy is equivalent to a two-step learning procedure with first step estimation satisfying desired econometric properties and characterize the unique linear equilibrium. Expected equilibrium prices exhibit underreaction. If the trading frequency is arbitrarily high, it is as if the market maker fully learns and reveals the unknown

distribution to the public. By introducing econometric tools to construct the process of getting unknown parameters, we analyze this strategic trading model that is robust to the ambiguity of a specific parameter.

CHAPTER II

Bounds for Treatment Effects in the Presence of Anticipatory Behavior

II.1 Introduction

This paper accommodates the matter of anticipation in the analysis of treatment effects by employing a potential outcomes framework (see e.g. Neyman (1923), Rubin (1974), Rubin (1978)) in a difference-in-differences model. The concept of anticipation is not unfamiliar to researchers in economics and social sciences, as seen in the work of Malani and Reif (2015), as well as Bošković and Nøstbakken (2018), for example. When anticipation occurs, forward-looking units change their behavior in reaction to the possibility of a new policy, and thus a treatment has an impact before its implementation. Therefore, it is crucial to consider the role of anticipation when evaluating an economic process and its outcome. However, despite its importance, most available published studies do not formally consider anticipation. A 'no anticipation' assumption is made, combined with a procedure of dropping data closely before the treatment if it is possibly violated, based on the argument that anticipation occurs only within a fixed time period prior to the introduction of a policy. Even in the few cases where anticipation is taken into account, the anticipatory behavior is accounted for in a restricted manner, such as an *ad hoc* restriction on

This chapter is based on the working paper "Bounds for Treatment Effects in the Presence of Anticipatory Behavior"(Gong (2021))

units' forward-looking behavior like rational or adaptive expectations.

When anticipatory behavior takes place, the identification strategies commonly used with multiple periods, such as the difference-in-differences model, fall apart. Consider an early retirement incentive program for teachers nearing the age of retirement. If those teachers foresee the possibility of retiring early, their behavior might change before the program is introduced. Due to the effect of future treatment status on pre-treatment periods, the observable pre-treatment outcomes are no longer drawn from the distribution of potential outcomes, if the treatment never takes place. Individuals will change their responses according to how they expect to be treated in the future. Thus, further information about units' anticipatory behavior is required. But such information is usually unattainable, since it is generally impossible to observe.

This paper provides novel strategies to build identified sets for treatment effects under assumptions restricting anticipatory behavior. Easy-to-implement estimation and inference strategies are also provided. I start from a difference-in-differences model with two time periods, and then generalize it to incorporate more complex models. I provide conditions for partial identification results of causal parameters when the anticipation status is unknown and incorporate anticipation in many widely used empirical designs. Employing a potential outcome framework, I analyze the treatment and the effects of anticipation based on the treatment rules, the anticipation assignments, and outcomes.

The departure from point identification starts with formulating restrictions on anticipatory behavior. In most cases, I do not have additional information, such as proxy variables, that helps us identify which participants have anticipated the policy change. As a result, I can say nothing about the pre-treatment distortion caused by anticipation. In this paper, I introduce a two-period difference-in-differences model where anticipation occurs in the first period and the treatment occurs in the second period. Further, I introduce two natural assumptions to help construct bounds for the treatment effect in the absence of such additional information. The first is a bound for the proportion of anticipators within the treatment group. This bound should be available from observed data. It can be a constant, or a parameter that

can be estimated. This is common practice in the literature, such as the work of Manski and Pepper (2013). The selection of this bound can vary from application to application, with one possible example being the treatment ratio. I provide models to motivate specific choices of the bound under various circumstances. The second assumption restricts the magnitude of the anticipatory effect. It requires that the absolute value of the anticipatory effect is no larger than that of the actual treatment effect. By doing so I build a link between the magnitude of an anticipator's reaction and the response to the implementation of the policy. Based on this an inequality between the average pre-treatment bias caused by anticipation and the treatment effect can be constructed with the help of the proportion of anticipators discussed above. Therefore, I can find a corresponding treatment effect range for the anticipators by characterizing how they react and linking that anticipatory effect to the actual treatment effect. Under these two assumptions, the fraction of units that anticipate the policy change may vary, but the average distortion caused by anticipation is successfully bounded and the parameter of interest is set identified.

As for the implementation purpose, I propose estimation and inference strategies based on easy-to-implement modifications to existing methods. The identification strategy provides an identified set with perfectly correlated and proportional upper and lower bounds. I propose a uniformly valid confidence set for my estimators with some modifications to Imbens and Manski (2004) under this specific setup. In their method, the upper and lower bounds of the confidence set are found by extending both sides of the identified set. The extension lengths are proportional to the standard errors of the bound estimators, and differ between upper and lower bounds. However, suppose this method is applied directly here. In that case, it may run into a counterintuitive situation where the confidence set for the treatment effect is shorter when the parameter is partially identified than when it is point-identified. I propose modifying this approach by extending both sides of the identified set by the same length proportional to the larger standard error of the two, which is a natural way to ensure the uniform validity. Analyzing this confidence set also provides researchers with further empirical implications. When the treatment and anticipatory effect go in different directions, I find a specific range of t-statistics

when the null hypothesis is zero treatment effect. If the t-statistic obtained when anticipation is ignored falls within this range, the conclusion of whether rejecting the null hypothesis does not change when considering anticipation. This confidence set also helps to build a framework for sensitivity analysis on certain conclusions of interest by choosing different bounds for anticipation possibility.

I apply the results of this paper to examine the effect of an early retirement incentive program on student achievement. This program is aimed at teachers near the age of retirement, and offers them financial incentives to retire before becoming eligible for full pension benefits. If the program is anticipated, eligible teachers may react in advance of its introduction, and such behavior might affect students' grades. The empirical results illustrate the potential pitfalls of failing to consider anticipation in program evaluation: the effect can be greatly overestimated in the worst case. I also conduct a sensitivity check by analyzing the level of anticipation probability one is willing to tolerate while maintaining the consistency of the original conclusion. It shows the conclusion is robust even when about three fourths of target units anticipate.

To permit the incorporation of anticipation in other common empirical setups, I provide several modifications. Instead of only focusing on the pre-treatment effect of anticipatory behavior in control group, I discuss the anticipatory behavior in control group by introducing an imperfect anticipation setup, where individuals make mistakes while anticipating. Post-treatment effects of anticipatory behavior are also discussed. To be consistent with common empirical approaches, generalizations to include covariates, multiple periods and nonlinear potential outcomes are provided and analyzed in the appendix.

This paper contributes to the literature on causal inference and program evaluation (Abadie and Cattaneo (2018), Athey and Imbens (2017)). Among them, my paper is most closely related to the work of Malani and Reif (2015), which discusses anticipatory behavior by interpreting the pre-trend phenomenon as a result of anticipation. Its authors propose a parametric time series model in which anticipation is an expectation of the future treatment for everybody by relying on the rational or adaptive expectation assumption. I incorporate the idea of anticipation in

a difference-in-differences framework with potential outcomes to remove parametric restrictions and allow heterogeneous anticipatory behavior among units. Heckman and Navarro (2007), under a different scenario, present a reduced form dynamic treatment effect model that also permits anticipation, but at the price of imposing further assumptions on the functional structure of the outcome equation.

This paper also contributes to the literature on difference-in-differences and event-study designs by considering anticipatory behavior. The additional anticipation could have impact prior to the introduction of a policy. Therefore, the present research is related to the literature aiming at more robust inference and identification strategies that allow for non-parallel trends assumptions, and to papers focusing on pre-trend analysis.

To interpret and deal with observed changes in outcomes prior to a treatment, Manski and Pepper (2018) propose a result on partial identification for the average treatment effect under "bounded variation" assumptions. This relaxes the parallel trends assumption by allowing for differences within a certain magnitude. Roth and Rambachan (2020) follow the idea that pre-treatment differences in trends are informative about counterfactual post-treatment differences and provide identification and inference results based on several common restrictions of this relationship. Freyaldenhoven, Hansen, and Shapiro (2019) propose a method that includes an additional covariate that is correlated with the outcomes through confounds only, and not treatments. Ye, Keele, Hasegawa, and Small (2021) propose a partial identification method for treatment effects with two groups of control units whose outcomes exhibit a negative correlation relative to the treated units. In this paper I interpret pre-trends as a result of unobservable anticipation activities. It is possible that people change their behavior because of their anticipation of future treatment. If people have information and may benefit by acting on it before a treatment, anticipation is a reasonable explanation for an observed pre-treatment effect, even when the parallel trends assumption is valid.

This paper is also complementary to the causal interpretation of event study coefficients, see Borusyak and Jaravel (2017), Sun and Abraham (2020), De Chaisemartin and d'Haultfoeuille (2020), and Goodman-Bacon (2021). With a generalization to

the longitudinal data, this paper can be regarded as relaxation of the 'no anticipation' assumption of these papers. This paper is also related more generally to the partial identification literature. In the present study, partial identification is obtained through moment inequalities, a method that is discussed in the handbook chapter of Molinari (2020). The sign restriction I impose is popular in time series literature to set-identify structural vector autoregressions (SVARs); see Kilian and Lütkepohl (2017).

The rest of this paper is organized as follows: Section II.2 generalizes the commonly used difference-in-differences model and introduces the basic setup about anticipation; Section II.3 provides extra assumptions and shows readers how to build the identified sets; Section II.4 describes estimation and inference; Section II.5 provides an empirical application; Section II.6 makes further discussion; and Section II.7 concludes. The mathematical proofs, together with some additional results, discussions and generalizations are collected in the supplemental appendix.

II.2 Setup and Assumptions

To illustrate anticipation in program evaluation, I consider an early retirement incentive program available for teachers, which offers experienced teachers financial incentives to retire before they would be eligible for full pension benefits. Suppose that one is interested in the effect of this early retirement incentive program on students' grades. There are several reasons to expect anticipation from teachers, whether treated or not, in this program. Teachers who anticipate may have received inside information from others, and it is also possible for them to speculate based on changes that have already happened. Younger teachers ineligible for the program won't react to it regardless of the anticipation status in both cases. However, teachers who anticipate the program and decide to retire early may put in less effort than younger teachers. Such behavior may harm students' grades before the implementation of the early retirement incentive program, and ignoring anticipation can lead to a bias while analyzing the effect of this program. The fact that teachers can anticipate based on unobservable information and adjust their behavior accordingly to

gain benefits implies that the future treatment will have an effect before its adoption and distort the treatment effect estimation if ignored. An accurate assessment of anticipation is therefore essential for the program evaluation.

II.2.1 The Basic Difference-In-Differences Model

To start with, I briefly describe the 'canonical' two-period difference-in-differences model in this section. As a well-understood starting point, this simple setting serves as a good baseline for understanding the approach I am going to use.

Consider a model with two periods $t \in \{0, 1\}$ and n units, $i \in \{1, \ldots, n\}$. Each unit is assigned an observable binary treatment D_i that takes value $d \in \{0, 1\}$ in the second period. The key identifying assumption requires that the treated and control group should change following parallel trends in the absence of treatment and the parameter of interest is the average treatment effect for treated (ATT).

Potential outcomes, defined below, depend on the time period and binary treatment status. The potential outcome for unit i in period t is denoted by the random variable $Y_{it}(d)$. Given a value of the implemented treatment d, the observed outcome of unit i at period t, Y_{it} can be written as

$$Y_{it} = \sum_{d \in \{0,1\}} Y_{it}(d)\mathbb{I}(D_i = d) = D_i Y_{it}(1) + (1 - D_i)Y_{it}(0)$$

and the parameter of interest $\mu = \mathbb{E}[Y_{i1}(1) - Y_{i1}(0)|D_i = 1]$. For identifying purpose, I need to assume **"Parallel Trends"** and **"No Anticipation"** which require

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0)|D_i = 0]$$

and

$$Y_{i0}(0) = Y_{i0}(1) \qquad \text{for all } i$$

Under these two assumptions, the parameter of interest μ is identified and for estimation purpose, I need to further put independent restrictions on the sampling process. The key idea here is that following the parallel trends assumption, one can

use the change in the control group to mimic that in the treated group if there were no treatment and get information about the unobservable potential outcomes for treated group in the absence of treatment in the post-treatment period. However, as pointed out, this approach requires no anticipatory behavior which states that the future treatment status should have no impact on pre-treatment outcomes. This assumption may be too restrictive in some situations, for example, the early retirement incentive program mentioned above. Thus it is important to figure out a way to accommodate anticipatory behavior under this difference-in-differences setup.

II.2.2 Introducing Anticipatory Behavior

In order to deal with the unobservable anticipatory behavior, I introduce another indicator for anticipation status. Suppose that in the first period, each unit has an unobservable binary anticipation status A_i that takes value $a \in \{0, 1\}$. Here $a = 1$ means this unit anticipates the future, and $a = 0$ means this unit does not anticipate the future. The potential outcomes can now be written as $Y_{it}(a, d)$. Introducing a second index in the expression of potential outcomes is commonly used when analyzing indirect effects, for example, analysis of spillover effects in Vazquez-Bare (2021). By implicitly assuming perfect anticipation, which means the anticipated treatment status should be the same as the actual treatment, I only focus on the pre-treatment anticipatory behavior in the treated group at this time. Anticipatory behavior in the control group and the post-treatment effect of the anticipatory behavior will be discussed later. After introducing another index for the anticipatory behavior, potential outcomes, defined below, can now depend on the binary treatment and anticipation status. I refer to the existence of the latent treatment in the first period as *anticipation* and the effect of the anticipatory behavior on unit i's potential outcome before the treatment occurs as the *anticipatory effect*.

As stated above, I am only focusing on the pre-treatment anticipatory behavior in the treated group now which means I am allowing $Y_{i0}(0, 1)$ and $Y_{i0}(1, 1)$ to be different from each other with no changes on $Y_{i0}(0)$, $Y_{i1}(0)$ and $Y_{i1}(1)$. With a little abuse of the notation, I am using the single index expression $Y_{it}(d)$ for the potential outcomes

when there is no difference in the value of anticipation status. With the existence of anticipatory behavior, the observed pre-treatment outcomes of the treated group are now a mixture of those who anticipate and those who don't. This brought extra difficulties in identification as one cannot tell the anticipators from those who do not anticipate and further assumptions are needed. To start with, I consider those assumptions that come from the canonical difference-in-differences model.

Assumption II.2.1 (Sampling)**.** $\{Y_{i0}(0,1), Y_{i0}(1,1), Y_{i0}(0), Y_{i1}(0), Y_{i1}(1), D_i, A_i\}_{i=1}^n$ *are independently and identically distributed across* i*.*

Assumption II.2.1 models the sampling process and states the potential outcomes, treatments, and unobservable anticipation status to be independent and identically distributed across units so that expectations are not indexed by i.

Assumption II.2.2. $Y_{i0}(0) = Y_{i0}(0,1)$.

Recall that $Y_{i0}(0)$ represents the pre-treatment potential outcome if one will not get treated where anticipation does not make a difference right now. One can understand this either as no anticipatory behavior happens or as a statement that people who will not get treated anticipate the future that they will not get treated perfectly so that they will not make any change. This implicit assumption can be relaxed later and I will make further discussion about it. Based on the above argument, assumption II.2.2 states that anticipation of a future treatment is the only channel through which future events affect the present. In the early retirement incentive program example, this assumption implies that the pre-treatment grades for students taught by the same teacher are indifferent regardless of the teacher's decision about early retirement if he has no anticipation of the program.

I can now make a table for potential outcomes in the difference-in-differences model with two periods.

	t=0	t=1
$D_i = 0$	$Y_{i0}(0)$	$Y_{i1}(0)$
$A_i = 0,\ D_i = 1$	$Y_{i0}(0,1)$	$Y_{i1}(1)$
$A_i = 1,\ D_i = 1$	$Y_{i0}(1,1)$	$Y_{i1}(1)$

Compared with the commonly used difference-in-differences framework, the critical difference is that the observed pre-treatment outcome Y_{i0} for the treated group is a mixture of the potential outcomes for those who do not anticipate $Y_{i0}(0,1)$ and those who anticipate $Y_{i0}(1,1)$ in the treated group in the pre-treatment period. $\mathbb{E}[Y_{i0}|D_i = 1]$ is no longer a good measure for the first period potential outcome without treatment for the treated group.

Under this setup, the parameter of interest I focus on is still the average treatment effect for treated (ATT) with a slight modification.

$$\mu_g = \mathbb{E}[g(Y_{i1}(1)) - g(Y_{i1}(0))|D_i = 1].$$

where $g(.)$ is a known measurable real function with $\mathbb{E}|g(Y)| < \infty$. Define the corresponding anticipatory effect for anticipators as

$$\tau_g = \mathbb{E}[g(Y_{i0}(1,1)) - g(Y_{i0}(0,1))|D_i = 1, A_i = 1].$$

The $g(.)$ function is slightly generalized from the commonly defined average treatment effect for treated (ATT). When $g(.)$ is the identity function, μ_g is the widely used ATT. If $g(.)$ is an indicator function like $g_u(Y) = \mathbb{I}(Y \leq u)$, then μ_g can be interpreted as the change in the probability for the outcomes to be no more than a specific cutoff u and can be used to help identify the distribution of potential outcomes. Different choices of this $g(.)$ function lead to different estimators. Introducing the g function helps handle some nonlinear structures for parameters I am interested in. For the simplicity of notation, I will write μ_g as μ and τ_g as τ when $g(.)$ is the identity function.

Assumption II.2.3 (Parallel Trends)**.**

$$\mathbb{E}[g(Y_{i1}(0)) - g(Y_{i0}(0))|D_i = 1] \quad = \quad \mathbb{E}[g(Y_{i1}(0)) - g(Y_{i0}(0))|D_i = 0].$$

Although the expression seems to be the same as the parallel trends assumption in the canonical difference-in-differences model, assumption II.2.3 requires that the treatment and control group change following parallel trends before and after the

treatment in the absence of both anticipation and the treatment.

Remark II.1. As explained above, for the benchmark model and basic results, I am only focusing on the pre-treatment anticipatory behavior in the treated group. For the anticipatory behavior in the control group, one can understand this simplification as an implicit assumption that all people anticipate their future correctly so that if one anticipates no future treatment, then he or she has no incentive to change his or her behavior. Whether one anticipates the future and realizes no treatment or no anticipation will not make any difference. Further modifications to incorporate anticipatory behavior in the control group and the post-treatment effect of anticipatory behavior will be discussed later.

Then I briefly discuss what the commonly used difference-in-differences estimator estimates without considering anticipation and compare it with the parameter that I am interested in. Throughout this part, I will choose $g(.)$ to be the identity function.

In the difference-in-differences regression model with two periods

$$Y_{it} = \beta_0 + \beta_1 t + \beta_2 D_i + \beta_3 t D_i + \varepsilon_{it}$$

Under Assumptions II.2.1-II.2.3, the coefficient of interest, β_3, can be written as

$$\begin{aligned}
\beta_3 &= \mathbb{E}[Y_{i1}|D_i = 1] - \mathbb{E}[Y_{i0}|D_i = 1] - \mathbb{E}[Y_{i1}|D_i = 0] + \mathbb{E}[Y_{i0}|D_i = 0] \\
&= \mu - \mathbb{P}[A_i = 1|D_i = 1]\tau \\
&= \mathbb{P}[A_i = 1|D_i = 1](\mu - \tau) + (1 - \mathbb{P}[A_i = 1|D_i = 1])\mu
\end{aligned}$$

If the difference-in-differences estimator is used directly, it will suffer from a bias equal to the average distortion caused by anticipation. This bias arises from the fact that the observable pre-treatment outcomes for the treated group do not reflect the potential outcomes for them without the treatment. Those who anticipate have already reacted in the first period and deviated from the parallel-trends benchmark. Thus, applying the difference-in-differences estimator directly suffers from a bias determined by both the proportion of those who anticipate and the magnitude of anticipatory effects. The last equality points out that this parameter can also be

written as a weighted average of the treatment effect μ for those who do not anticipate and the net treatment effect after the adoption of the policy for those who do anticipate $\mu - \tau$. In general, the relationship between this estimand and the treatment effect depends on the sign of the anticipatory effects. Suppose the treatment and anticipatory effects have the same sign. In that case, anticipation will drive the difference-in-differences estimator towards zero relative to the treatment effect because of contamination. The idea of this distortion is captured in the graphs below. Figure II.1 shows the result obtained by applying the difference-in-differences estimator directly while Figure II.2 describes the situation that considers anticipation. The distortion between β_3 and μ is caused by anticipation.

II.3 Upper and Lower Bounds for Treatment Effects

II.3.1 Main Results

Anticipation makes the commonly used difference-in-differences estimator a mixture of anticipatory and treatment effects. The fundamental difficulty in obtaining identification is distinguishing between those who anticipate and those who do not. This section introduces several assumptions to build upper and lower bounds for treatment effects under different circumstances. Motivations for specific assumptions are provided. The following results link observed outcomes, potential outcomes, treatment assignments, and anticipation status and are used in further discussions.

From the analysis above, it is clear that two unobservable variables are contributing to the pre-treatment distortion. One is the possibility for treated units to anticipate, $\mathbb{P}[A_i = 1|D_i = 1]$, and the other is the anticipatory effect for anticipators τ_g. These two variables both need to be analyzed to recover the treatment effect. If there is a reasonable proxy available for the anticipation treatment, then one can use this proxy variable to measure the anticipation status for each unit. However, such a proxy variable is not always available. To overcome the difficulty of not being able to tell people who anticipate from others, I introduce a bounding parameter $\pi \in (0, 1)$ that summarizes how the anticipation probability can be bounded. Here

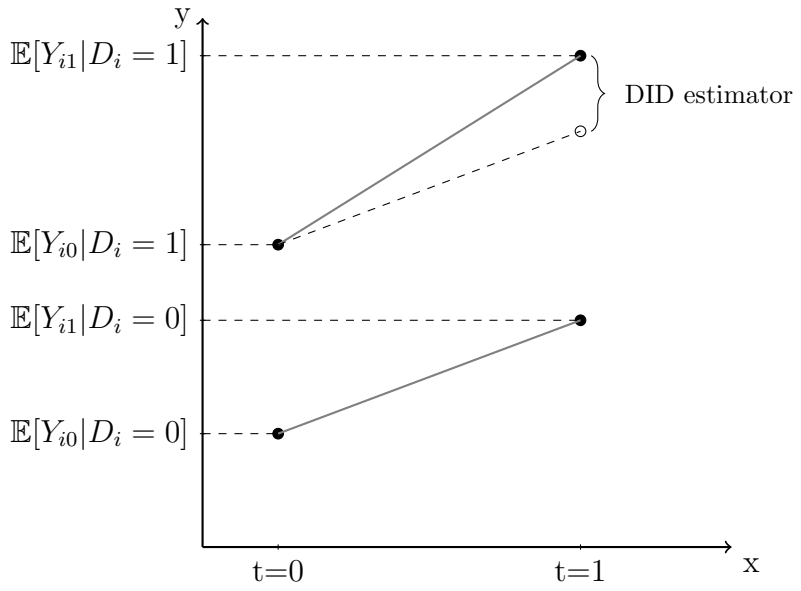

Figure II.1: Difference-In-Differences without Anticipation

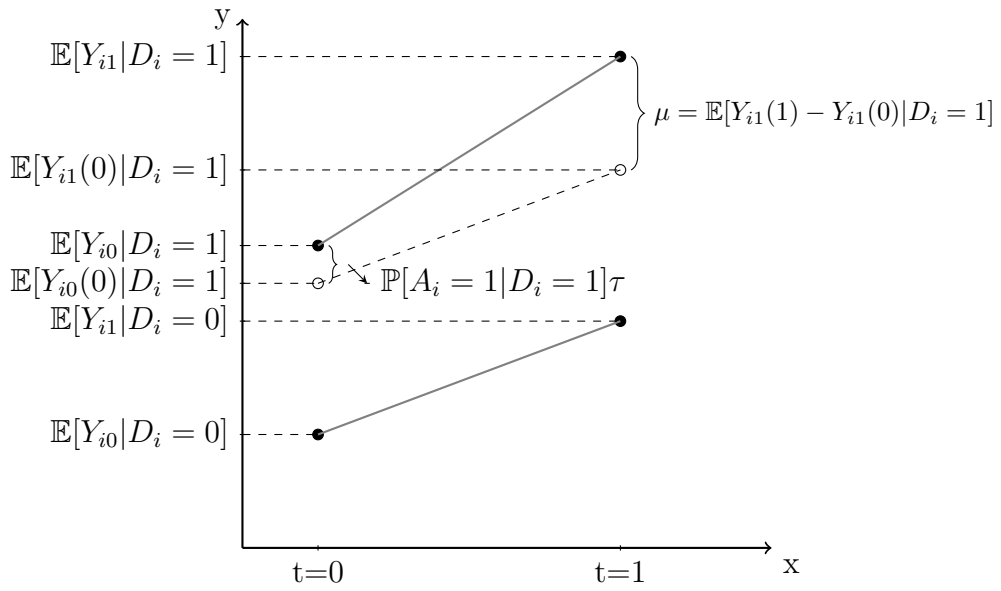

Figure II.2: Difference-In-Differences With Anticipation

π is a parameter that can be obtained from available information, including the treatment assignment and outcomes of units, and it can be either constant or at least estimated from observable terms. Researchers can choose a π based on their empirical setups, and I will discuss possible choices of π in the next section.

Assumption II.3.1. $\mathbb{P}[A_i = 1|D_i = 1] \leq \pi$

For the anticipatory effect τ_g, although I cannot observe it directly, I can build a relationship between it and the treatment effect for treated μ_g. τ_g is caused by people's anticipation of a possible future treatment and behavior before the treatment to gain benefit. People's reactions and behavior are guided by their own guesses of the future policy. On the other hand, μ_g measures the treatment effect treated people receive when the policy is adopted. This effect happens based on revealed policy and treatment status. For example, suppose someone is going to sell his property at a lower than usual price because of anticipation of possible negative price shock in the future. In that case there is no reason for him to accept a price that is even lower than the price when the shock really comes. Therefore, it is reasonable to expect that the magnitude of treatment effects should be no smaller than that of the anticipatory effect as the former is a reaction based on known information while the latter one is based on uncertain information. Even when they anticipate the future, they may be reluctant to it. In the example of the early retirement incentive program, this statement requires that the effect caused by teachers who put less effort because of anticipating a possible early retirement opportunity is no larger than the treatment effect when the early retirement incentive program is implemented. Furthermore, as the anticipatory effect and treatment effect may not be in the same direction, I only impose restrictions on the magnitude.

Assumption II.3.2. $|\tau_g| \leq |\mu_g|$

Assumption II.3.1 and Assumption II.3.2 help build bounds for the two unobservable terms, the proportion of anticipators among treated and the magnitude of anticipatory effect separately and based on the above assumptions, the parameter of interest μ_g is partially identified using observed variables, especially with the help of the commonly used difference-in-differences estimand.

Theorem II.1. *Under Assumptions II.2.1-II.3.2, the parameter of interest μ_g is partially identified via a closed interval in the following form.*

$$\mu_g \in m_g \left[\min \left\{ 1, \frac{1}{1 - sgn(\tau_g \mu_g)\pi} \right\}, \max \left\{ 1, \frac{1}{1 - sgn(\tau_g \mu_g)\pi} \right\} \right]$$

with $m_g = \mathbb{E}[g(Y_{i1}) - g(Y_{i0})|D_i = 1] - \mathbb{E}[g(Y_{i1}) - g(Y_{i0})|D_i = 0]$.

Theorem II.1 points out that the treatment effect is located in an interval where the difference-in-differences parameter without anticipation is one of its bounds. The other bound is obtained by enlarging or reducing it by a specific ratio depending on the bounding parameter π and signs of treatment and anticipation effects. As shown in Figures II.1 and II.2, the distortion only happens within the group of treated and anticipate units, so once the sign of the anticipatory effect is determined, the sign of the bias is also determined. The difference-in-differences parameter without anticipation is by design one side of the interval. If anticipatory behavior happens in both the control and the treated group, then distortions happen in both groups, and the sign of the bias is ambiguous. The distortion bias has a limited magnitude restricted by both the bounding parameter π and the treatment effect magnitude, so I can build partial identification results for the parameter of interest based on observables.

Although I impose the bounds for anticipation probability and magnitude restrictions, these assumptions are not the only way to build partial identification results for the parameter of interest with anticipation. There might be empirical setups where these assumptions are not reasonable, and researchers would like to impose alternative assumptions, such as bounded outcomes or further conditional independent restrictions. These assumptions are also reasonable under specific situations, such as when the $g(.)$ function I am interested in is bounded by itself. It is not the case that identified set under one assumption is tighter than the other so I should pick one of them, but different sets of assumptions may be reasonable under different empirical circumstances. Incorporating more combinations of alternative assumptions and providing identification results allow us to incorporate anticipation in more situations and give researchers the freedom to modify assumptions based on

the empirical setup. I propose several different combinations of assumptions as well as corresponding upper and lower bounds expression of the treatment effects in the appendix.

II.3.2 Choice of π

This section discusses several possible choices of the bounding parameter π for the anticipation probability among treated units under different setups.

Example II.1 $\pi = \pi_0$ where π_0 is a constant number. If π is a constant number, this implies that there is a common upper bound for the possibility of anticipation. This choice of π may be consistent with the setup where people get treated randomly receive private information that helps with anticipation. Then the overall anticipation possibility should be no more than the proportion of people that have access to this private information.

Example II.2 $\pi = \mathbb{P}[D_i = 1]$. This example states that the possibility for people within the treatment group to anticipate does not exceed the proportion of people treated at last. This argument follows the idea that anticipation will happen when a future treatment sends some signals and unobservable information in advance. These are the bases for someone to anticipate a treatment. Suppose the density of these signals caused by future adoptions of policies is related to the overall scope of the treatment. In that case, it is reasonable to use the treated probability to help bound the proportion of people who anticipate. I will explain this choice and corresponding assumptions later in a model where people anticipate from public information.

Example II.3 The univariate bound can be modified to incorporate the idea of stratification. Suppose researchers are willing to divide units into several subgroups and allow anticipation behavior to differ among subgroups. In that case, I can pick π as a k dimensional vector if I have k subgroups in total. For instance, the vector of assignment can be summarized according to genders or geographical areas, and researchers can get a bound separately for each subgroup. This can also be regarded as a bound conditional on a discrete variable that divides the group based on several categories and link to the case with covariates.

Example II.4 Suppose anticipation behavior happens among known reference groups for each unit, as mentioned in Manski (2013). In that case, I can pick π based on subgroup information. For example, if researchers would like to use the treatment ratio to capture the density of information and on the other hand they also believe this kind of interaction only happens among units within a specific geographical distance, they can pick π as the treatment ratio for each subgroup defined by the given geographical distance.

Besides, the choice of π can also play the role of sensitivity analysis. The expression of the bounds should be monotonic in π, and researchers can use different choices of π to explore the robustness of obtained conclusions by checking the specific cutoff under which the consistency of conclusion can be maintained. This sensitivity analysis also helps us understand to what extent the conclusion depends on the choice of bounds, and researchers can report the range of anticipation probability that rejects a particular null hypothesis.

II.3.2.1 A Toy Economic Behavior Model

This section provides a toy economic behavior model that motivates the choice of $\pi = \mathbb{P}[D_i = 1]$ and explains what assumptions one needs to imply this choice of bounding parameter. Consider a setup where a future policy has led to some public information prior to its implementation, and people can take advantage of this kind of signals to anticipate. For example, there might be rumors and changes in teaching assignments before the early retirement incentive program occurs and teachers may anticipate based on them. The density of the signals is correlated with the overall treatment ratio as a higher ratio of experienced teachers that are eligible for the program implies more people are interested in the policy, and they may talk more about it. There will be more information about this program.

For simplicity, consider a case where $\mathbb{P}[A_i = 1|D_i = 0] = \mathbb{P}[A_i = 1|D_i = 1]$, which means the probability to anticipate is the same between the control and treatment groups. Suppose the density of information generated by the future implementation of the treatment, denoted by u, is proportional to the treatment ratio $\mathbb{P}[D_i = 1]$ in

the form

$$u = \alpha \mathbb{P}[D_i = 1] \qquad \alpha > 0,$$

and that this information is known to both treated and control groups. This formula captures the idea that a higher treatment ratio will lead to a case where the information needed for anticipation is more explicit for people.

For each unit i, I introduce a random variable U_i that represents the level of information to which one needs to be exposed for anticipation to occur. A higher U_i means this unit needs more signals to realize the possible treatment. In contrast, a lower U_i indicates that this unit has a keen observation and can reach a conclusion with less information. I will use $F(.)$ to represent the c.d.f of U_i across the population. For any single unit i, the anticipatory behavior follows

$$A_i = \mathbb{I}[U_i \leq u],$$

which means the density of information needs to be larger than the cutoff U_i for unit i to anticipate, and I have

$$\mathbb{P}[A_i = 1] = \mathbb{P}[U_i \leq u] = F(\alpha \mathbb{P}[D_i = 1]).$$

I further assume that $f'(.) \geq 0$ where $f(.)$ is the probability density function of the random variable U_i. This assumption states that the fraction of people who can marginally anticipate increases with the level of information needed to form anticipation, consistent with the intuition that people who can anticipate based on little information should be small.

Based on the assumptions above, I can conclude that

$$\frac{\partial^2 \mathbb{P}[A_i = 1]}{\partial \mathbb{P}[D_i = 1]^2} = \alpha^2 f'(\alpha \mathbb{P}[D_i = 1]) \geq 0,$$

which means $\mathbb{P}[A_i = 1]$ should be a convex function on interval $[0, 1]$ with respect to $\mathbb{P}[D_i = 1]$. If $\mathbb{P}[D_i = 1] = 0$, I conclude $\mathbb{P}[A_i = 1] = 0$ as there is nothing to anticipate. I also know that when $\mathbb{P}[D_i = 1] = 1$, $\mathbb{P}[A_i = 1] \leq 1$. This combined

with the convex function argument shows

$$\mathbb{P}[A_i = 1] \leq \mathbb{P}[D_i = 1] \times 1 + 0 = \mathbb{P}[D_i = 1].$$

The above model illustrates that under a setup where units can get information public among all the groups but unobservable to econometricians and anticipate future treatments, Assumption II.3.1 will be satisfied by choosing $\pi = \mathbb{P}[D_i = 1]$ as long as one is willing to assume fewer people can anticipate from less information and as information accumulates, the number of people who marginally learn it increases.

II.4 Estimation and Inference

The previous section illustrates that by using a difference-in-differences approach, the treatment effect for treated with anticipation is partially identified under certain assumptions. The population average expressions of the interval bounds lead to straightforward estimators using sample means under independent assumptions. This section builds uniformly effective confidence sets for the partially identified parameters.

Assume that researchers observe data from a distribution $\mathrm{P} \in \mathbf{P}$ with the unobservable parameter, $\mathbb{P}[A_i = 1|D_i = 1] \in [0, \pi]$. $\mathbf{P}$ refers to the family of distributions that satisfy the sampling, potential outcomes restrictions. For the inferential goal under partial identification, I would like to build a confidence set that is uniformly consistent in level α i.e.

$$\lim_{n\to\infty} \inf_{\mathrm{P}\in\mathbf{P}, \mathbb{P}[A_i=1|D_i=1]\in[0,\pi]} \mathbb{P}[\mu_g \in CS^{\mu}_{\alpha}] \geq \alpha,$$

where CS^{μ}_{α} is the α level confidence set for the parameter of interest μ_g.

Here I provide confidence sets based on Imbens and Manski (2004), Stoye (2009), and Stoye (2020). The upper and lower bounds for the identified set are estimated using the same sample and thus highly correlated. I can build an easy-to-implement confidence set by modifying the method addressed above. For notation simplicity,

call the upper and lower bounds of parameter μ_g as $\mu_{g,u}$ and $\mu_{g,l}$. A uniformly effective confidence set can be built if the corresponding estimators $\hat{\mu}_{g,u}$ and $\hat{\mu}_{g,l}$ exist and satisfy

Assumption II.4.1. $\sqrt{n}\begin{pmatrix} \hat{\mu}_{g,l} - \mu_{g,l} \\ \hat{\mu}_{g,u} - \mu_{g,u} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_l^2 & \rho\sigma_l\sigma_u \\ \rho\sigma_l\sigma_u & \sigma_u^2 \end{bmatrix}\right)$ *uniformly in* $P \in \mathbf{P}$*, and there are estimators* $(\hat{\sigma}_l^2, \hat{\sigma}_u^2, \hat{\rho})$ *converge to their population values uniformly in* $P \in \mathbf{P}$

Assumption II.4.2. *For all* $P \in \mathbf{P}$, $\underline{\sigma}^2 \leq \sigma_l^2, \sigma_u^2 \leq \bar{\sigma}^2$ *for some positive and finite* $\underline{\sigma}^2$ *and* $\bar{\sigma}^2$ *and* $\mu_{g,u} - \mu_{g,l} = \Delta \leq \bar{\Delta} < \infty$

Theorem II.2. *Under Assumption II.4.1 and II.4.2, define* $\hat{\sigma} = \max\{\hat{\sigma}_l, \hat{\sigma}_u\}$ *and find* C_n *that satisfies*

$$\Phi\left(C_n + \sqrt{n}\frac{\hat{\mu}_{g,u} - \hat{\mu}_{g,l}}{\hat{\sigma}}\right) - \Phi(-C_n) = \alpha.$$

Φ *represents the cumulative distribution function for standard normal distribution. Then I have*

$$\lim_{n\to\infty} \inf_{P\in\mathbf{P}, \mathbb{P}[A_i=1|D_i=1]\in[0,\pi]} \mathbb{P}\left(\mu_g \in \left[\hat{\mu}_{g,l} - C_n\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu}_{g,u} + C_n\frac{\hat{\sigma}}{\sqrt{n}}\right]\right) \geq \alpha.$$

To be consistent with the setup in the empirical application, I focus on the case $\tau_g \leq 0 \leq \mu_g$ as an example, and I have $\mu_g \in m_g\left[\frac{1}{1+\pi}, 1\right]$. Corresponding bound estimators will be

$$\begin{aligned} \hat{\mu}_{g,u} &= \frac{1}{n_1}\sum_{i=1}^{n}[g(Y_{i1}) - g(Y_{i0})]D_i - \frac{1}{n_0}\sum_{i=1}^{n}[g(Y_{i1}) - g(Y_{i0})](1 - D_i) \\ \hat{\mu}_{g,l} &= \frac{\hat{\mu}_{g,u}}{1 + \hat{\pi}} \qquad n_1 = \sum_{i=1}^{n} D_i \qquad n_0 = n - n_1 \qquad \hat{\pi} \text{ is a consistent estimator for } \pi \end{aligned}$$

If one would like to pick $\pi = \mathbb{P}[D_i = 1]$ then a straightforward $\hat{\pi}$ will be $\frac{1}{n}\sum_{i=1}^{n} D_i$. The standard errors can be found in the supplemental appendix.

The assumptions and results mainly follow Imbens and Manski (2004) and Stoye (2009). When compared with the Imbens and Manski (2004) approach, the confidence set I construct has a slight difference that I choose to extend along with the upper and lower bounds by the same length $C_n\frac{\hat{\sigma}}{\sqrt{n}}$, while in Imbens and Manski (2004) the same critical value is picked but the standard errors are different. An intuitive explanation is that although the estimators of the upper and lower bounds are ordered by construction, which is an important assumption mentioned in Stoye (2009), the upper and lower bounds can have reverse order and the interval changes from $[\frac{m_g}{1+\pi}, m_g]$ to $[m_g, \frac{m_g}{1+\pi}]$ when the confidence set contains both positive and negative values. Therefore the corresponding variances for the estimators of upper and lower bound need to be accommodated to use the larger one for both bounds. This modification works for the construction of confidence sets with perfectly correlated and proportional upper and lower bounds, especially when the confidence set contains zero. The proof is discussed in the appendix.

The change in the expression of confidence sets changes the significance level of rejecting specific null hypothesis, $H_0 : \mu_g = 0$ in many cases compared with the situation without anticipation. One interesting case worth mentioning happens when μ_g and τ_g have different signs. For the specific null hypothesis $H_0 : \mu_g = 0$, I can calculate the value of t-statistics that guarantee the conclusion of whether rejecting it or not unchanged regardless of anticipation.

Corollary II.1. *Suppose t^* satisfies*

$$\Phi(t^*) - \Phi(-t^*/2) = \alpha.$$

Φ represents the cumulative distribution function for standard normal distribution and the inferential goal is to test $H_0 : \mu_g = 0$ at level α. Suppose confidence set CS^{μ}_{α} is constructed following the procedure in Theorem II.2. If μ_g and τ_g have different signs and the t-statistic from the difference-in-differences model without anticipation $\tilde{t}$ satisfies $|\tilde{t}| > t^*$, then for any π, $0 \notin CS^{\mu}_{\alpha}$.

Corollary II.1 gives empirical researchers a specific cutoff t^* for the most common case of testing $H_0 : \mu_g = 0$. If the absolute value of t-statistic exceeds t^* for the

case of different signs, then taking anticipation into consideration will not change the conclusion of rejecting the null hypothesis. For example, when $\alpha = 0.95$, the corresponding t^* is 3.3. This means if the absolute value of t-statistic is larger than 3.3 without anticipation, you can still reject zero hypothesis for treatment effect regardless of the anticipation probability when the treatment and anticipatory effects have different signs. This corollary gives empirical researchers a cutoff where they can claim the effectiveness of their conclusions even with anticipation as long as the t-statistics is large enough.

II.5 Empirical Application

In this section, I illustrate the results of this paper in the environment established by Fitzpatrick and Lovenheim (2014), which analyzes the effects of an early retirement incentive program on students' achievement. The authors conducted a difference-in-differences based analysis using exogenous variations from the early retirement incentive (ERI) program targeting on teachers in Illinois during the mid-1990s to evaluate the effect of large-scale teacher retirements on student achievement. The Teacher Retirement System in Illinois requires retired members who are at least 55 years old and have 20 years of service experience to collect pension benefits at a 6% discount rate below age 60. If both the employer and employee pay a one-time fee, an Early Retirement Option allows eligible members to collect their full benefit. In 1992-1993 and 1993-1994, an early retirement incentive (ERI) program was offered as an alternative to ERO, which allowed employees to buy five extra years of age and experience as long as they retired immediately, and this allowed those with at least 50 years old and 15 years of service credit to increase their retirement benefits.

Notably, the ERI programs may impact students' learning as this might lead to a change in teachers' experience and age structure, which will eventually influence students' grades. In their paper, the authors used a difference-in-differences approach to analyze how promoting the ERI program affected the students' grades. It turned out that they found no evidence of an adverse effect and even a positive effect on grades in some circumstances. I analyze the average treatment effect for treated by

taking anticipation into consideration. The outcome of interest is students' grades, and the major difference is now teachers might anticipate the program in advance and make benefit from it.

The authors collected data from several sources. Teacher Service Record is an administrative dataset that contains information of employees from Illinois Public Schools. The second set of data provides school-level information on test scores for given subjects and grades. The third source contains demographic information of students in schools. The analysis is restricted to teachers of third, sixth and eighth grades as standardized testing in Illinois focuses on these grades. One major issue for the data is the ERI take-up is not observed directly. The authors exploited the fact that teachers with 15 or more years of experience were most likely to take up the program and used it as a proxy for the intensity of treatment by the ERI program.

Consider the restrictions I impose on potential outcomes for the case with anticipation. It requires that the students' grades of teachers that are ineligible or choose not to retire early should not be affected, and it also requires if teachers are not aware of this program in advance, there should be no change in the grades. Further, it restricts once the ERI program is implemented, whether teachers anticipate or not should no longer affect the students' grades. The parallel trends assumption requires that trends in students' grades among schools with fewer treated teachers are precise counterfactuals for trends among schools with more treated teachers without anticipation.

Following the idea that teachers, whether eligible or not, may have some information from a third party before the implementation of the ERI program so that they may have anticipated something, I pick $\pi = \mathbb{P}[D_i = 1]$ and the probability of getting treated is estimated by calculating the proportion of experienced teachers with more than 15 years of service credit and I get correpsonding $\hat{\pi}$. Recall that this bound is used to capture the intensity of potential unobservable information, which is proportional to the intensity of treatment, and I can use the proportion of teachers with more than 15 years of teaching experience to bound the anticipation probability. The magnitude effect assumption requires that the anticipatory effect, τ, which is the result of potential behavior changes of teachers who think it is possible for

them to retire early, has a smaller magnitude compared with the treatment effect, μ, which is the change in students grades caused by the ERI program after its implementation. Even under a perfect anticipation setup from econometricians' view, the teachers themselves do not have the confidence that they would definitely be eligible for the policy and take it up so it is reasonable to expect the anticipatory effect does not have a larger magnitude than the treatment effect when the policy occurs. Further, it is reasonable to expect the treatment effect μ to have the same sign as the non-negative difference-in-differences estimator from Fitzpatrick and Lovenheim (2014). On the other hand, I follow the argument in the same paper that claims teachers near the retirement age and anticipate the possibility of early retirement may put less effort than younger teachers. Therefore, it is reasonable to argue that the anticipatory effect τ is non-positive.

With all the assumptions discussed above, I can analyze the treatment effect with anticipation starting from the following equation in Fitzpatrick and Lovenheim (2014) that estimates the difference-in-differences estimator.

$$Y^s_{igt} = \beta_0 + \beta_1 (Teachers \geq 15)_{ig} \times Post_t + \beta_2 Teachers_{ig} \times Post_t + \gamma \boldsymbol{X}_{it} + \delta_{ig} + \varphi_{tg} + \varepsilon^s_{itg}$$

Y^s_{igt} is the test score of grade g for subject s in school i and year t. $Teachers \geq 15$ is the number of teachers with at least 15 years of experience before 1994 and thus eligible for the program. $Teachers$ is the average total numbers of teachers. $Post$ serves as the period, an indicator variable that equals one after the school year of 1993. The vector $\boldsymbol{X}$ contains demographic information while δ and φ are corresponding fixed effect terms. Although covariates are included here, the parametric assumption that it enters the outcome linearly implies that the treatment effect is homogeneous across different values of controls. The intensity of information related to the choice of π has already been captured by the proportion of experienced teachers. If there is no anticipation, β_1 from this equation estimates the effect of the ERI program on students' grades. Based on the estimator for β_1 and π I choose, I can analyze the results with anticipation. I check the results for different grades and subjects and compare the cases for all teachers. Results are shown in Table II.1. Similar results

using data from subject-specific teachers are listed in the supplemental appendix.

Table II.1: Effects of the Early Retirement Incentive Program on Test Scores

	Original Results		With Anticipation	
	Math	Reading	Math	Reading
All Grade	0.003	0.009	[0.002,0.003]	[0.006,0.009]
	(0.004)	(0.003)		
	[-0.003,0.010]	[0.002,0.015]	[-0.004,0.010]	[0.001,0.014]
Grade 3	0.002	-0.009	[0.001,0.002]	[-0.018,-0.009]
	(0.01)	(0.008)		
	[-0.017,0.021]	[-0.025,0.008]	[-0.018,0.021]	[-0.05,0.022]
Grade 6	-0.0001	0.006	[-0.0002,-0.0001]	[0.004,0.006]
	(0.005)	(0.004)		
	[-0.01,0.01]	[-0.003,0.015]	[-0.022,0.021]	[-0.004,0.014]
Grade 8	0.005	0.013	[0.003,0.005]	[0.008,0.013]
	(0.005)	(0.005)		
	[-0.005,0.015]	[0.004,0.022]	[-0.006,0.014]	[0.001,0.021]

Notes: This table contains data for all teachers. Each column presents results from a separate regression. Teachers who teach multiple grades are included in each grade. Teachers who teach in self-contained classrooms are assumed to teach both math and English. I list identified sets in the first row and 95% level confidence sets in the third row for each result with anticipation. For comparison purposes, I also provide estimators, standard errors and 95% confidence intervals for results from Fitzpatrick and Lovenheim (2014). Standard errors are displayed with parentheses.

For the partial identification results, I provide identified sets as well as the 95% confidence sets. I notice that the estimate is at times negative from the initial results in Fitzpatrick and Lovenheim (2014). As these negative estimates are insignificant at the 95% level, I conclude that this distortion error is due to the finite sample bias. For these estimators, I adjust my way to get the identified sets and confidence sets by changing the sign restriction and find that the confidence sets, after incorporating anticipation, still cannot reject the null hypothesis $\mu = 0$ no matter which sign I choose. The changes in the result are mainly in two aspects. On the one hand, the results with anticipation suggest the treatment effect can be smaller than the one

we get directly from the difference-in-differences approach as the DID estimator also captures the pre-treatment negative effect caused by anticipation. The effect can be overestimated up to about 30% because of anticipation. On the other hand, the confidence sets, compared with the difference-in-differences approach, are slightly shifted leftwards and this result also reminds people to be more careful when interpreting the non-negative treatment effect. Despite these differences, results incorporating anticipation still support the conclusion that the ERI programs have a non-negative effect on students' grades. These results imply that incorporating anticipation can make the result more robust and still support our idea of the non-negative effect of ERI programs on student achievement.

I conduct a robustness check to see the range of choices for π that keeps the significance of the estimator at a 95% level and show the result in Figure II.3. I focus on the effect of the early retirement incentive program on the reading grade in grade 8. I present the identified set as well as the 95% confidence set for a sequence of π, including 0.1, 0.25, 0.5, $\mathbb{P}[D_i = 1]$, 0.75 and 0.9. The shorter interval represents the identified set while the longer one represents the confidence set. I observe that at an anticipation probability of 0.75, more precisely around 0.7, the confidence set marginally contains point 0, which means this positive treatment effect is quite robust even when taking anticipation into consideration. The null hypothesis will only be rejected when about three fourths of the target teachers anticipate it.

II.6 Discussion

This section discusses modifications on the two-period difference-in-differences model that only considers the pre-treatment anticipatory behavior in the treated group to incorporate anticipation in broader setups. These generalizations build the anticipation framework on more empirical related assumptions and cover problems researchers encounter in applied work.

First we focus on the restrctions on pre-treatment anticipatory behavior in the treated group only. In our discussion about the Assumption II.2.2, I said that one can understand the focus only on anticipatory behavior within the treated group

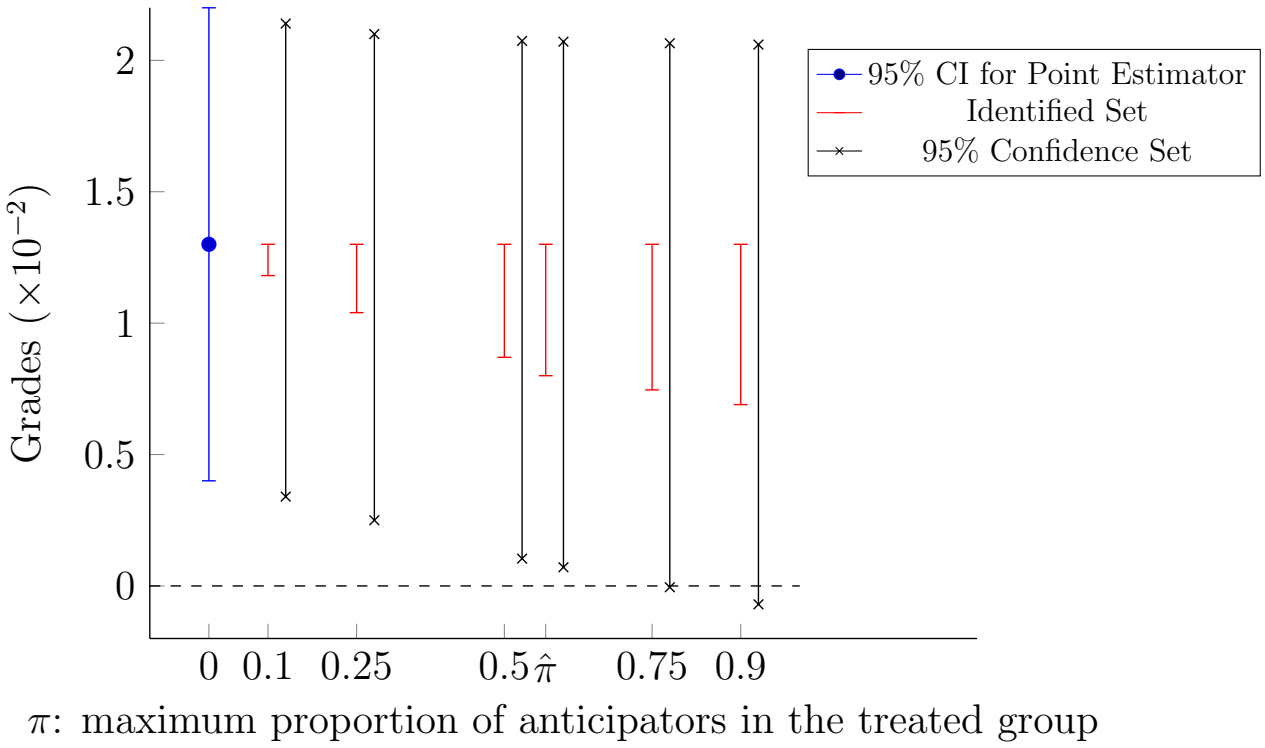

Figure II.3: Robustness Analysis

as an implicitly assumption of 'perfect anticipation', which implies that units that make anticipation will get the anticipated treatment status in the future. However, this might be too strong in some circumstances. For example, sometimes people may anticipate the existence of a specific policy but they are not clear whether or not they will get treated. In the early retirement incentive program example, it is possible for teachers to anticipate the possibility of early retirement but they are not sure about the exact amount of service credit they can buy and thus cannot anticipate their future treatment status perfectly. I consider the consequences if a mistake is made when anticipating a future treatment in this section and successfully incorporate the anticipatory behavior in the control group. Furthermore, it is essential to explore the robustness of our conclusion by checking the error rate under which consistent conclusions can still be obtained. If researchers aim to test a particular null hypothesis, they can also report the lowest error rate at which the null hypothesis is no longer rejected.

In order to distinguish between anticipated treatment status and the actual treatment units receive and incorporate imperfect anticipation, I now define the anticipation status variable A_i as a random variable that takes three values $\{-1, 0, 1\}$. The difference between $A_i = 0$ and $A_i \neq 0$ distinguishes between those who anticipate and those who don't. However, among those who anticipate, $A_i = 1$ indicates that this unit makes correct anticipation while $A_i = -1$ indicates a wrong one. The potential outcome for unit i in period t still depends on both anticipation and treatment (a, d) and is denoted by the random variable $Y_{it}(a, d)$. The only difference is now $a \in \{-1, 0, 1\}$ and A_i is no longer a binary treatment. Compared with the benchmark model, some modifications need to be made on the assumptions to incorporate imperfect anticipation.

The random sampling assumption remains unchanged and the only modification is that anticipatory behavior also happens within the control group so we have more potential outcomes now and we need another index in the expression of pre-treatment potential outcomes for the control group as well.

Assumption II.6.1. $\{Y_{i0}(a, d), Y_{i1}(d), D_i, A_i\}_{i=1}^n$ *are independently and identically distributed across* i*.*

The assumption that requires anticipation is the only channel for the future to affect present is still needed. However, it needs to be addressed that now one's pre-treatment behavior is affected by his anticipated status, which is likely to be different from the treatment status he receives in the future.

Assumption II.6.2. *The potential outcomes satisfy*

$$Y_{i0}(0,0) = Y_{i0}(0,1) = Y_{i0}(1,0) = Y_{i0}(-1,1) \qquad Y_{i0}(-1,0) = Y_{i0}(1,1)$$

Assumption II.6.2 mainly describes two groups of units. The first group either does not anticipate or they anticipate they will not get treated in the future so they will behave in the same way. People in the second group expect that they will get treated in the future and they will behave in the other way. As anticipation is the only way I am allowing the future to affect present, one's pre-treatment behavior is decided by their anticipated treatment status. A treated person with the wrong anticipation should have the same anticipated treatment status as an untreated person who anticipates correctly and thus they should behave in the same way as those who do not anticipate. However, those who will get treated and anticipate correctly should behave in the same way as those who won't be treated but anticipate wrongly as they all think they will be covered. This assumption points out that what drives people's pre-treatment behavior is their beliefs of the treatment status. Trying to distinguish between anticipated treatment status and real treatment received is important in the case where people make mistakes while anticipating.

One may argue that in the situation of imperfect anticipation, it is possible that people no longer make clear anticipation about the future treatment and they believe they will get treated at a probability. Different people hold different beliefs about their treatment possibilities and behave differently. This situation can be regarded as a case where anticipation is a multivalue treatment and people with different beliefs receive different levels of anticipation treatment. As all the things related to anticipation are unobservable, introducing more levels of different anticipation treatments also requires more assumptions regarding each group. Therefore I still focus on the case where people's anticipation about the future is whether he is treated

or not.

The parameter of interest μ_g is still

$$\mu_g = \mathbb{E}[g(Y_{i1}(1)) - g(Y_{i1}(0))|D_i = 1]$$

with similar restrictions on $g(.)$ function. As people's post treatment outcomes are not affected by anticipation status so I still use one index to represent the potential outcomes $Y_{it}(d)$. The anticipatory effect is modified as

$$\begin{aligned} \tau_g &= \mathbb{E}[g(Y_{i0}(1,1)) - g(Y_{i0}(0,0))|D_i = 1, A_i = 1] \\ &= \mathbb{E}[g(Y_{i0}(-1,0)) - g(Y_{i0}(0,0))|D_i = 1, A_i = -1] \end{aligned}$$

where I implicitly require that the anticipatory effect for those who anticipate correctly and wrongly are the same.

Assumption II.6.3.

$$\mathbb{E}[g(Y_{i1}(0)) - g(Y_{i0}(0,0))|D_i = 1] = \mathbb{E}[g(Y_{i1}(0)) - g(Y_{i0}(0,0))|D_i = 0].$$

The key idea of the parallel trend is to require those who get treated and not get treated should behave in the same way without the treatment. Following this idea, I need to pick those who have an anticipated untreated status when compared with the outcome in the first period.

Assumption II.6.4.

$$\mathbb{P}[A_i \neq 0|D_i = 1], \mathbb{P}[A_i \neq 0|D_i = 0] \leq \pi$$

$$\mathbb{P}[A_i = -1|D_i = 1, A_i \neq 0] = \mathbb{P}[A_i = -1|D_i = 0, A_i \neq 0] = \varepsilon.$$

The first part of Assumption II.6.4 requires us to pick a π as the bound for the probability of anticipation among treated and control groups. This is straightforward as under the 'perfect anticipation' situation, units that won't get treated will not react

to the anticipation but now they may react to it because of the wrong anticipation. If one would like to argue there is a specific relationship between the possibility of anticipation within treated and control groups, this assumption might be relaxed. For now I am assuming a common bound π for two probabilities is picked. In the second part, I assume the fraction of units that make wrong anticipation is known as ε across treated and control groups. Recall that in the discussion about the bias caused by anticipation, I point out that the bias is driven by those who anticipate and react to it in the first period. Under the setup of imperfect anticipation, the proportion of units that cause the bias is determined by anticipated treatment status and thus related to both the proportion of those who anticipate and the accurate rate among anticipators.

Assumption II.6.5. $|\tau_g| \leq |\mu_g|$.

Assumption II.6.5 is the same magnitude restriction as before. Based on the assumptions above, now we are able to partially identify the parameter of interest μ_g.

Theorem II.3. *Under Assumptions II.6.1-II.6.5, the parameter of interest μ_g is partially identified via a closed interval based on $m_g = \mathbb{E}[g(Y_{i1}) - g(Y_{i0})|D_i = 1] - \mathbb{E}[g(Y_{i1}) - g(Y_{i0})|D_i = 0]$, π and ε. Define $\mu_{g,1}(\varepsilon) = \frac{m_g}{1+sgn(\tau_g\mu_g)\pi\varepsilon}$ and $\mu_{g,2}(\varepsilon) = \frac{m_g}{1-sgn(\tau_g\mu_g)\pi(1-\varepsilon)}$. Then we have form*

$$\mu_g \in [\min\{\mu_{g,1}(\varepsilon), \mu_{g,2}(\varepsilon)\}, \max\{\mu_{g,1}(\varepsilon), \mu_{g,2}(\varepsilon)\}]$$

If $\varepsilon = 0$, which represents the situation of perfect anticipation, this interval degenerates to the interval we get in the benchmark model for the treatment effect. Compared with the interval I get for the treatment effect above, one significant difference is now the treatment effect is not bounded by the difference-in-differences estimator from one side. That difference derives from the fact that both the control group and the treated group deviate in the first period. For a perfect anticipation setup, those in the control group will not react to the anticipation and only the treated group is moving either upward or downward depending on the signs of antic-

ipatory effect. When imperfect anticipation is allowed and both control and treated groups react to it, the distortion in the first period can be either positive or negative depending on the anticipation possibility in different groups. To accommodate this framework in more empirical settings, I don't impose specific assumptions on the relationship between the possibility to anticipate in different groups. If in specific situations, for example, a case where those who get treated receive private information, researchers are comfortable to decide the relationship between these two possibilities, it is possible to improve the identified set based on further assumptions.

Although I start with the case where ε is known, a better explanation for including the error rate of anticipation is to understand this procedure as a sensitivity check. Researchers can pick different possible error rates ε and analyze the region where their conclusions are robust to the pick of error rate. Further, if a specific null hypothesis is tested, the error rate among which the conclusion holds consistently can also be reported. This helps people to understand to what extent the conclusion is affected by the assumption of perfect anticipation.

Another thing that one might be interested in is, what if the anticipatory behavior in the treated group has an effect on the post treatment behavior and the post-treatment potential outcomes in the treated group are also different between those who anticipate and those who do not. Starting from the benchmark model, now let us assume that $Y_{i1}(0,1)$ and $Y_{i1}(1,1)$ are different. Then we have two effects related to anticipation $\tau_1 = \mathbb{E}[g(Y_{i0}(1,1)) - g(Y_{i0}(0,1))|D_i = 1]$ and $\tau_2 = \mathbb{E}[g(Y_{i1}(1,1)) - g(Y_{i1}(0,1))|D_i = 1]$. Following similar logic above, one can find that $\mu_g = m_g + \mathbb{P}[A_i = 1|D_i = 1](\tau_1 - \tau_2)$, which implies that if no more assumptions about the relationship between pre and post treatment effect caused by anticipation are imposed then nothing more can be said. Further, if one would like to assume that the effect caused by anticipation remains unchanged before and after the treatment, this equation points out that the existence of anticipation will have no effect on the identification and estimation of the parameter of interest here. This is a generalization of the no anticipation assumption in the canonical difference-in-differences model where both effects are assumed to be zero.

Besides the modifications mentioned above, further generalizations including the

model incorporates anticipation with covariates in multiple periods as well as non-linear outcomes that involves change-in-changes model are also provided in the appendix.

II.7 Conclusion

This paper proposes a potential outcome framework for analyzing treatment effects with the presence of anticipatory behavior. Based on a two-period difference-in-differences model, the findings of this paper show how the standard estimator can be biased and provide a weighted average of the treatment and the anticipatory effect. I also provide conditions under which I can obtain upper and lower bounds for the treatment effects. The motivation and implication of each assumption are discussed to accommodate empirical research backgrounds. This paper contributes to the empirical research by introducing anticipation in a practical and easy to generalize way starting from the classical difference-in-differences model, which makes it robust to the existence of this kind of forward looking behavior. An easy-to-implement estimation and inference strategy is also provided. I propose a sensitivity analysis approach based on it that discusses the validity of conclusions under different restrictions on the anticipation possibility, and this approach suggests a specific range of t-statistics that guarantee the effectiveness of the conclusion got without anticipation when the treatment and anticipatory effect have different signs. I illustrate the results in this paper by examining the effect of early retirement incentive programs on student achievement while considering anticipation and show potential pitfalls if anticipation is ignored. To make this framework more general and less restrictive, I provide several modifications based on the two-period difference-in-differences model to be consistent with common empirical setups.

The analysis for this paper still leaves some open questions, with some of them discussed in the appendix. I provide several alternative combinations of assumptions that can be used to obtain partial identification results for treatment effects with anticipation. For example, bounded outcomes assumptions and further conditional independence restrictions on potential outcomes and treatments. The choice

of these assumptions depends on the empirical backgrounds researchers are working on. Alternative assumptions combined with available bounds provide applied workers with more choices that fit into broad applied circumstances. Further work can focus on some frequent issues in empirical studies, for example, anticipation effects related to instrumental variables. When a time gap exists between the instrumental variable and the treatment, the instrumental variable can cause people to anticipate future treatment and thus react to it before the treatment occurs. This setup is also related to cases with imperfect compliance and situations where anticipation will affect people's future selections into treatments. Another possible extension is to incorporate the anticipation phenomenon in the synthetic control framework. It makes sense as the treatments in typical synthetic control applications are often big policy changes that would naturally be anticipated. Following Ferman and Pinto (2019) the anticipation treatment can be regarded as an unobservable confounder that is correlated with treatment because only those get treated in the future will react to anticipation. In that case, the pre-treatment weight that fits well may not construct good counterfactual post-treatment outcomes for the treated unit and thus causes problem. Trying to analyze the behavior of synthetic control estimator and difference-in-difference estimator with anticipation and compare the performance of these two approaches will be of interest for applied work. By taking these situations into consideration, I am more likely to incorporate anticipation in more diverse empirically relevant situations and introduce it to more applied models.

CHAPTER III

Cluster Robust Inference in Linear Regression Models with Many Covariates

III.1 Introduction

In empirical research, it is common practice to safeguard against possible correlation between specific units by employing the cluster-robust standard error. At the same time, researchers often include a lot of control variables in the linear model to control for confounders. However, the inclusion of a large set of control variables can be problematic for the commonly used inference procedure, even without the cluster structure, see Cattaneo et al. (2018) and Jochmans (2020), for example. When introducing clustering structure, this problem can be more severe and needs further adjustment.

Motivated by the observations above, this paper studies the consequences of allowing the error term to be correlated within clusters with a high-dimensional covariates. To be more specific, the dimension of the covariates are allowed, but not required to be high and as we are focusing on the OLS-based inference procedures, we are not allowing the covariates have a higher dimension than the sample size.

Our main purpose is to build a valid inference procedure for the OLS-based estimator that is robust to the existence of clustering and the existence of many

This chapter is based on the working paper "Cluster Robust Inference in Linear Regression Model with Many Covariates" with Matias D. Cattaneo, Michael Jansson and Whitney K. Newey .

covariates with as least restrictions on the cluster designs and the covariate structure as possible. We get several main results there. First of all, we provide sufficient conditions to guarantee the asymptotic normality of the OLS-based estimator under this setup. Second, we analyze a class of variance estimators that incorporate several commonly used cluster robust standard errors and analyze the performance of these estimators. On one hand, we point out the conditions under which some previously used standard errors may still be effective with a trade-off between the dimension of covariates and the largest cluster sample size. On the other hand, we also propose an alternative cluster robust standard error that is consistent under such setup and more flexible cluster designs. Another separate conclusion that is consistent with what we get in Cattaneo et al. (2018) in heteroskedasticity case shows that the jackknife estimator, a cluster generalization of the HC-3 estimator is conservative even when we have many covariates, which means if one reject the null hypothesis using the jackknife estimator, he or she should be confident about the conclusion.

On one hand, our paper contributes to the huge literature on cluster robust standard errors in linear models. There are a series of papers about cluster robust inference that varies from empirical guide to specific inference problems in clustering designs, for example, "few" clusters problem. See, Bertrand, Duflo, and Mullainathan (2004), Cameron and Miller (2015), Ibragimov and Müller (2016), Abadie, Athey, Imbens, and Wooldridge (2017), Conley, Gonçalves, and Hansen (2018), MacKinnon (2019), Hansen and Lee (2019), Esarey and Menger (2019), Canay, Santos, and Shaikh (2021), MacKinnon, Nielsen, and Webb (2022), for example. Our paper contributes to this literature by analyzing a new class of cluster robust standard error under the setup of containing high dimensional covariates in the linear model and allowing as much flexibility as possible in cluster designs.

On the other hand, our paper also adds to the literature in linear models whose number of regressors is non-ignorable compared with sample size. To be more precise, the condition shows that the dimension of covariates matters when the largest cluster size combined with the increasing speed of the covariates' dimension compared with the sample size is non-ignorable. The main method generalizes from Cattaneo et al. (2018) and also discusses results related to Jochmans (2020). Generalizations from

independent error terms to cluster design is not trivial in the sense that when the clusters are allowed to contain infinite elements within each of them, the convergence result is affected not only by the dimension of covariates, but also by the cluster size. Further, the difference in the correlation structures between covariates and error terms are also allowed to bring more flexibilities. D'Adamo (2019) also puts an eye on this problem, however, the results in that paper only focus on the case where each cluster has finite number of elements with some extra assumptions restricting the behavior of the asymptotic variance directly. That paper also does not study the property of the estimator by providing further conditions to guarantee the validity.

The rest of this article is organized as follows. Section 2 introduces the basic setup and main assumptions. Section 3 discusses the class of standard errors we study and provides a general expression for all the standard errors mentioned here. Section 4 gives the main results of the article. Section 5 reports the simulation results and section 6 concludes. Proofs as well as additional theoretical results are discussed in the appendix.

III.2 Setup and Assumptions

Suppose $\{(y_{i,n}, \mathbf{x}'_{i,n}, \mathbf{w}'_{i,n}) : 1 \leq i \leq n\}$ is generated by a model of the form

$$y_{i,n} = \boldsymbol{\beta}' \mathbf{x}_{i,n} + \boldsymbol{\gamma}'_n \mathbf{w}_{i,n} + u_{i,n}, \qquad i = 1 \dots n, \tag{III.1}$$

where $\mathbf{x}_{i,n}$ is of fixed dimension d and $\mathbf{w}_{i,n}$ is of (possibly) growing dimension K_n. Our main goal is to conduct valid OLS-based inference for the parameter $\boldsymbol{\beta}$ that is robust to clustering as well as the existence of high dimensional covariates. Here, the high dimension works in the sense that K_n cannot be a complete ignorable part compared with the sample size.

Similarly to Cattaneo et al. (2018), henceforth CJN, we impose three high-level conditions. To state the first condition, let $\mathcal{X}_n$ and $\mathcal{W}_n$ denote collections of random

variables satisfying $\mathbb{E}[\mathbf{x}_{i,n}|\mathcal{X}_n] = \mathbf{x}_{i,n}$ and $\mathbb{E}[\mathbf{w}_{i,n}|\mathcal{W}_n] = \mathbf{w}_{i,n}$, respectively, and define

$$U_{i,n} = y_{i,n} - \mathbb{E}[y_{i,n}|\mathcal{X}_n, \mathcal{W}_n] \qquad \text{and} \qquad \mathbf{V}_{i,n} = \mathbf{x}_{i,n} - \mathbb{E}[\mathbf{x}_{i,n}|\mathcal{W}_n].$$

Also, let the cardinality of a set A be denoted $\#A$.

Assumption III.2.1. *Assume that* $\mathcal{C}_{\mathcal{S},n} = \max_{1\leq G\leq N_{\mathcal{S},n}} \#\mathcal{S}_{G,n} = o(\sqrt{n})$ *and* $\mathcal{C}_{\mathcal{T},n} = \max_{1\leq g\leq N_{\mathcal{T},n}} \#\mathcal{T}_{g,n} = o(\sqrt[3]{n})$, *where* $\{\mathcal{S}_{G,n} : 1 \leq G \leq N_{\mathcal{S},n}\}$ *and* $\{\mathcal{T}_{g,n} : 1 \leq g \leq N_{\mathcal{T},n}\}$ *are partitions of* $\{1, \ldots, n\}$ *such that* $\{\mathbf{V}_{s,n} : s \in \mathcal{S}_{G,n}\}$ *are independent over* G *conditional on* $\mathcal{W}_n$ *and* $\{U_{t,n} : t \in \mathcal{T}_{g,n}\}$ *are independent over* g *conditional on* $(\mathcal{X}_n, \mathcal{W}_n)$.

To state the next condition, define

$$\tilde{\mathbf{\Gamma}}_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{V}}_{i,n}\tilde{\mathbf{V}}_{i,n}',$$

where

$$\tilde{\mathbf{V}}_{i,n} = \sum_{j=1}^{n} M_{ij,n}\mathbf{V}_{j,n}, \qquad M_{ij,n} = \mathbb{I}(i=j) - \mathbf{w}_{i,n}'\left(\sum_{k=1}^{n}\mathbf{w}_{k,n}\mathbf{w}_{k,n}'\right)^{-1}\mathbf{w}_{j,n}.$$

Also, define

$$\mathbf{U}_n(g) = (U_{t_{g,n}(1),n}, \ldots, U_{t_{g,n}(\#\mathcal{T}_{g,n}),n})', \qquad g = 1, \ldots, N_{\mathcal{T},n},$$

where $t_{g,n}(\cdot)$ is any function such that $\{t_{g,n}(1), \ldots, t_{g,n}(\#\mathcal{T}_{g,n})\} = \mathcal{T}_{g,n}$. Finally, define

$$\begin{aligned}\mathcal{C}_n \;=\; & \max_{1\leq i\leq n}\{\mathbb{E}[U_{i,n}^4|\mathcal{X}_n, \mathcal{W}_n] + \mathbb{E}[\|\mathbf{V}_{i,n}\|^4|\mathcal{W}_n]\} \\ & + \max_{1\leq g\leq N_{\mathcal{T},n}}\{1/\lambda_{\min}(\mathbb{E}[\mathbf{U}_n(g)\mathbf{U}_n(g)'|\mathcal{X}_n, \mathcal{W}_n])\} + 1/\lambda_{\min}(\mathbb{E}[\tilde{\mathbf{\Gamma}}_n|\mathcal{W}_n]),\end{aligned}$$

where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of its argument.

Assumption III.2.2. $\mathbb{P}[\lambda_{\min}(\sum_{i=1}^{n}\mathbf{w}_{i,n}\mathbf{w}_{i,n}') > 0] \to 1$, $\limsup_{n\to\infty} K_n/n < 1$, *and* $\mathcal{C}_n = O_p(1)$.

To state the last condition, define

$$\begin{aligned}
\varrho_n &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[R_{i,n}^2], \qquad R_{i,n} = \mathbb{E}[u_{i,n}|\mathcal{X}_n, \mathcal{W}_n], \\
\rho_n &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[r_{i,n}^2], \qquad r_{i,n} = \mathbb{E}[u_{i,n}|\mathcal{W}_n], \\
\chi_n &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\mathbf{Q}_{i,n}\|^2], \qquad \mathbf{Q}_{i,n} = \mathbb{E}[\mathbf{v}_{i,n}|\mathcal{W}_n],
\end{aligned}$$

where $\|\cdot\|$ denotes the Euclidean norm and where

$$\mathbf{v}_{i,n} = \mathbf{x}_{i,n} - \mathbb{E}\left[\sum_{j=1}^{n}\mathbf{x}_{j,n}\mathbf{w}_{j,n}'\right]\left(\mathbb{E}\left[\sum_{j=1}^{n}\mathbf{w}_{j,n}\mathbf{w}_{j,n}'\right]\right)^{-1}\mathbf{w}_{i,n}$$

is the population counterpart of

$$\hat{\mathbf{v}}_{i,n} = \mathbf{x}_{i,n} - \left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_{j,n}\mathbf{w}_{j,n}'\right)\left(\frac{1}{n}\sum_{j=1}^{n}\mathbf{w}_{j,n}\mathbf{w}_{j,n}'\right)^{-1}\mathbf{w}_{i,n} = \sum_{j=1}^{n}M_{ij,n}\mathbf{x}_{j,n}.$$

Assumption III.2.3. *$\chi_n = O(1)$, $\varrho_n + n(\varrho_n - \rho_n) + n\chi_n\varrho_n = o(1)$, and $\frac{1}{n^2}\sum_{i=1}^{n}\|\hat{\mathbf{v}}_{i,n}\|^4 = o_p(1)$.*

Remark III.1.

- We invariably set $\mathcal{X}_n = (\mathbf{x}_{1,n}, \ldots, \mathbf{x}_{n,n})$, but it is convenient to allow $\mathcal{W}_n \neq (\mathbf{w}_{1,n}, \ldots, \mathbf{w}_{n,n})$.
- If Assumption 1 of CJN is satisfied, then Assumption III.2.1 is satisfied with $\{\mathcal{S}_{G,n}\} = \{\mathcal{T}_{g,n}\}$, $\mathcal{C}_{\mathcal{S},n} = \mathcal{C}_{\mathcal{T},n} = O(1)$, and $\mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n, \mathcal{W}_n] = 0$ for $i \neq j$. Our main objective is to relax the latter assumption, but we will also explore the consequences of allowing $\mathcal{C}_{\mathcal{T},n}$ to grow. When doing so, it turns out to be convenient to relax the requirement $\{\mathcal{S}_{G,n}\} = \{\mathcal{T}_{g,n}\}$.
- If $\mathbb{E}[U_{i,n}U_{j,n}|\mathcal{X}_n, \mathcal{W}_n] = 0$ for $i \neq j$, then

$$\max_{1\leq g\leq N_{\mathcal{T},n}}\{1/\lambda_{\min}(\mathbb{E}[\mathbf{U}_n(g)\mathbf{U}_n(g)'|\mathcal{X}_n, \mathcal{W}_n])\} = \max_{1\leq i\leq n}\{1/\mathbb{E}[U_{i,n}^2|\mathcal{X}_n, \mathcal{W}_n]\}.$$

As a consequence, if Assumption 1 of CJN is satisfied, then Assumption III.2.2 is equivalent to Assumption 2 of CJN.

- If $\max_{1\leq i\leq n}\mathbb{E}[\|\mathbf{V}_{i,n}\|^2|\mathcal{W}_n]=O_p(1)$ and if $\chi_n=O(1)$, then $n^{-1}\sum_{i=1}^n\|\hat{\mathbf{v}}_{i,n}\|^2=O_p(1)$. If also $n^{-1/2}\max_{1\leq i\leq n}\|\hat{\mathbf{v}}_{i,n}\|=o_p(1)$, then

$$\frac{1}{n^2}\sum_{i=1}^n\|\hat{\mathbf{v}}_{i,n}\|^4\leq\left(\frac{1}{\sqrt{n}}\max_{1\leq i\leq n}\|\hat{\mathbf{v}}_{i,n}\|\right)^2\left(\frac{1}{n}\sum_{i=1}^n\|\hat{\mathbf{v}}_{i,n}\|^2\right)=o_p(1).$$

As a consequence, if Assumption III.2.2 is satisfied, then Assumption III.2.3 is implied by Assumption 3 of CJN.

III.3 Variance Estimators

It is convenient to write the OLS estimator $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}+\left(\sum_{i=1}^n\hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}'\right)^{-1}\left(\sum_{i=1}^n\hat{\mathbf{v}}_{i,n}u_{i,n}\right)$$

and our first result provides the conditions under which the OLS estimator follows the asymptotic normality distribution with an infeasible estimator in the following form.

$$\boldsymbol{\Omega}_n^{-1/2}\sqrt{n}(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta})\to_d\mathcal{N}(\mathbf{0},\mathbf{I}_d),\qquad\boldsymbol{\Omega}_n=\hat{\boldsymbol{\Gamma}}_n^{-1}\boldsymbol{\Sigma}_n\hat{\boldsymbol{\Gamma}}_n^{-1},\tag{III.2}$$

where

$$\hat{\boldsymbol{\Gamma}}_n=\frac{1}{n}\sum_{1\leq i\leq n}\hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}'\qquad\text{and}\qquad\boldsymbol{\Sigma}_n=\frac{1}{n}\mathbb{V}\left[\sum_{1\leq i\leq n}\hat{\mathbf{v}}_{i,n}u_{i,n}\middle|\mathcal{X}_n,\mathcal{W}_n\right].$$

Theorem III.1. *Suppose Assumptions III.2.1-III.2.2 hold and suppose Assumption III.2.3 holds with*

$$\mathcal{C}_{\mathcal{S},n}\rho_n=o(1)\qquad\textit{and}\qquad\frac{\mathcal{C}_{\mathcal{T},n}^3}{n^2}\sum_{1\leq i\leq n}||\hat{\mathbf{v}}_{i,n}||^4=o_p(1).$$

Then (III.2) holds.

If (III.2) is satisfied and if $\mathbf{\Sigma}_n^{-1} = O_p(1)$, then a (variance) estimator $\hat{\mathbf{\Sigma}}_n$ satisfying $\hat{\mathbf{\Sigma}}_n = \mathbf{\Sigma}_n + o_p(1)$ will also satisfy

$$\hat{\mathbf{\Omega}}_n^{-1/2}\sqrt{n}(\hat{\beta}_n - \beta) \to_d \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \qquad \hat{\mathbf{\Omega}}_n = \hat{\mathbf{\Gamma}}_n^{-1}\hat{\mathbf{\Sigma}}_n\hat{\mathbf{\Gamma}}_n^{-1}. \tag{III.3}$$

Under Assumption III.2.1, the matrix $\mathbf{\Sigma}_n$ is given by

$$\mathbf{\Sigma}_n = \frac{1}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}} \hat{\mathbf{V}}_n(g)'\mathbb{E}[\mathbf{U}_n(g)\mathbf{U}_n(g)'|\mathcal{X}_n, \mathcal{W}_n]\hat{\mathbf{V}}_n(g),$$

where $\hat{\mathbf{V}}_n(g) = (\hat{\mathbf{v}}_{t_{g,n}(1),n}, \ldots, \hat{\mathbf{v}}_{t_{g,n}(\#\mathcal{T}_{g,n}),n})'$. For our purposes, it turns out to be convenient to work with the folllowing alternative representation, obtained using standard properties of the Kronecker product:

$$\mathbf{\Sigma}_n = \mathrm{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'\mathbb{E}[\mathbf{U}_n(g)\otimes\mathbf{U}_n(g)|\mathcal{X}_n, \mathcal{W}_n]\right\},$$

where vec_d^{-1} is the inverse of the vectorization operator $\mathrm{vec}_d : \mathbb{R}^{d\times d} \to \mathbb{R}^{d^2}$. In what follows, we consider estimators of $\mathbf{\Sigma}_n$ obtained by replacing each $\mathbb{E}[\mathbf{U}_n(g)\otimes\mathbf{U}_n(g)|\mathcal{X}_n, \mathcal{W}_n]$ with an estimator.

The simplest plausible estimator of $\mathbb{E}[\mathbf{U}_n(g)\otimes\mathbf{U}_n(g)|\mathcal{X}_n, \mathcal{W}_n]$ is arguably $\hat{\mathbf{u}}_n(g)\otimes\hat{\mathbf{u}}_n(g)$, where $\hat{\mathbf{u}}_n(g) = (\hat{u}_{t_{g,n}(1),n}, \ldots, \hat{u}_{t_{g,n}(\#\mathcal{T}_{g,n}),n})'$. The associated estimator of $\mathbf{\Sigma}_n$ is the Liang and Zeger (1986) estimator

$$\begin{aligned}\hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}} &= \frac{1}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}} \hat{\mathbf{V}}_n(g)'\hat{\mathbf{u}}_n(g)\hat{\mathbf{u}}_n(g)'\hat{\mathbf{V}}_n(g) \\ &= \mathrm{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'(\hat{\mathbf{u}}_n(g)\otimes\hat{\mathbf{u}}_n(g))\right\},\end{aligned}$$

a degrees-of-freedom corrected version of which (used in Stata) is

$$\check{\mathbf{\Sigma}}_n^{\mathtt{LZ}} = \mu_n \hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}}, \qquad \mu_n = \frac{N_{\mathcal{T},n}}{N_{\mathcal{T},n}-1}\frac{n-1}{n-K_n}.$$

In the special case where $\mathcal{C}_{\mathcal{T},n} = 1$ (i.e., when each $\mathcal{T}_{g,n}$ is a singleton), these estimators reduce to the so-called HC0 and HC1 estimators, respectively, and it follows from CJN that the estimators are inconsistent in general when $K_n/n \nrightarrow 0$. Also, because

$$\mu_n = \frac{1}{1-K_n/n}\{1+o(1)\}$$

when $\mathcal{C}_{\mathcal{T},n} = o(n)$, the estimators $\check{\mathbf{\Sigma}}_n^{\mathtt{LZ}}$ and $\hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}}$ are not asymptotically equivalent when $K_n/n \nrightarrow 0$. In fact, even if $K_n/n \to 0$ the estimators $\check{\mathbf{\Sigma}}_n^{\mathtt{LZ}}$ and $\hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}}$ can fail to be asymptotically equivalent because $\mathbf{\Sigma}_n = O_p(\mathcal{C}_{\mathcal{T},n}) \neq O(1)$ in general. On the other hand, suppose $\mathcal{C}_{\mathcal{T},n}\mathcal{M}_n = o_p(1)$, where

$$\mathcal{M}_n = 1 - \min_{1\leq i\leq n} M_{ii,n}.$$

Then $\check{\mathbf{\Sigma}}_n^{\mathtt{LZ}} = \mathbf{\Sigma}_n + o_p(1)$ whenever $\hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}} = \mathbf{\Sigma}_n + o_p(1)$, the reason being that $\mu_n\mathbf{\Sigma}_n = \mathbf{\Sigma}_n + o_p(1)$ because $\mathcal{M}_n \geq K_n/n$.

For $g, h \in \{1, \ldots, N_{\mathcal{T},n}\}$, let $\mathbf{M}_n(g,h)$ be a $(\#\mathcal{T}_{g,n}) \times (\#\mathcal{T}_{h,n})$ matrix obtained by partitioning

$$\mathbf{M}_n = \mathbf{I}_n - \mathbf{W}_n(\mathbf{W}_n'\mathbf{W}_n)^{-1}\mathbf{W}_n', \qquad \mathbf{W}_n = (\mathbf{W}_n(1)', \ldots, \mathbf{W}_n(N_{\mathcal{T},n})')',$$

as

$$\mathbf{M}_n = \begin{pmatrix} \mathbf{M}_n(1,1) & \cdots & \mathbf{M}_n(1,N_{\mathcal{T},n}) \\ \vdots & \ddots & \vdots \\ \mathbf{M}_n(N_{\mathcal{T},n},1) & \cdots & \mathbf{M}_n(N_{\mathcal{T},n},N_{\mathcal{T},n}) \end{pmatrix},$$

where $\mathbf{W}_n(g) = (\mathbf{w}_{t_{g,n}(1),n}, \ldots, \mathbf{w}_{t_{g,n}(\#\mathcal{T}_{g,n}),n})'$. Assuming each $\mathbf{M}_n(g,g)$ is invertible, a "bias reduced" (in the terminology of Imbens and Kolesar (2016)) estimator of $\mathbf{\Sigma}_n$

is given by

$$\hat{\mathbf{\Sigma}}_n^{\mathtt{BR}} = \mathrm{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'(\mathbf{M}_n(g,g)^{-1/2}\hat{\mathbf{u}}_n(g)\otimes\mathbf{M}_n(g,g)^{-1/2}\hat{\mathbf{u}}_n(g))\right\}.$$

By construction, this estimator reduces to CJN's version of HC2 when $\mathcal{C}_{\mathcal{T},n} = 1$. Similarly, an estimator that reduces to CJN's version of HC3 when $\mathcal{C}_{\mathcal{T},n} = 1$ is the "jackknife" estimator

$$\hat{\mathbf{\Sigma}}_n^{\mathtt{JK}} = \mathrm{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'(\mathbf{M}_n(g,g)^{-1}\hat{\mathbf{u}}_n(g)\otimes\mathbf{M}_n(g,g)^{-1}\hat{\mathbf{u}}_n(g))\right\}.$$

Although not necessarily consistent, this estimator turns out to be asymptotically conservative under weak conditions even when $K_n/n \nrightarrow 0$.

A cluster robust analog of the "Hadamard" estimator $\hat{\mathbf{\Sigma}}_n^{\mathtt{HC}}$ studied by CJN is given by

$$\hat{\mathbf{\Sigma}}_n^{\mathtt{CR}} = \mathrm{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'\kappa_n^{\mathtt{CR}}(g,h)(\hat{\mathbf{u}}_n(h)\otimes\hat{\mathbf{u}}_n(h))\right\},$$

where

$$\begin{aligned}&\begin{pmatrix}\kappa_n^{\mathtt{CR}}(1,1) & \cdots & \kappa_n^{\mathtt{CR}}(1,N_{\mathcal{T},n})\\ \vdots & \ddots & \vdots\\ \kappa_n^{\mathtt{CR}}(N_{\mathcal{T},n},1) & \cdots & \kappa_n^{\mathtt{CR}}(N_{\mathcal{T},n},N_{\mathcal{T},n})\end{pmatrix}\\ =&\begin{pmatrix}\mathbf{M}_n(1,1)\otimes\mathbf{M}_n(1,1) & \cdots & \mathbf{M}_n(1,N_{\mathcal{T},n})\otimes\mathbf{M}_n(1,N_{\mathcal{T},n})\\ \vdots & \ddots & \vdots\\ \mathbf{M}_n(N_{\mathcal{T},n},1)\otimes\mathbf{M}_n(N_{\mathcal{T},n},1) & \cdots & \mathbf{M}_n(N_{\mathcal{T},n},N_{\mathcal{T},n})\otimes\mathbf{M}_n(N_{\mathcal{T},n},N_{\mathcal{T},n})\end{pmatrix}^{-1}.\end{aligned}$$

By construction, this estimator reduces to CJN's $\hat{\mathbf{\Sigma}}_n^{\mathtt{HC}}$ when $\mathcal{C}_{\mathcal{T},n} = 1$. More importantly, this estimator turns out to be consistent under conditions permitting

$K_n/n \nrightarrow 0$.

The estimators $\hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}}, \check{\mathbf{\Sigma}}_n^{\mathtt{LZ}}, \hat{\mathbf{\Sigma}}_n^{\mathtt{BR}}, \hat{\mathbf{\Sigma}}_n^{\mathtt{JK}}$, and $\hat{\mathbf{\Sigma}}_n^{\mathtt{CR}}$ can be embedded in a class of estimators that can be analyzed in a unified way. To define this class, let $N_{\kappa,n} = \sum_{1\leq g\leq N_{\mathcal{T},n}}(\#\mathcal{T}_{g,n})^2$ and let κ_n be a symmetric $N_{\kappa,n}\times N_{\kappa,n}$ matrix partitioned as

$$\kappa_n = \begin{pmatrix} \kappa_n(1,1) & \cdots & \kappa_n(1,N_{\mathcal{T},n}) \\ \vdots & \ddots & \vdots \\ \kappa_n(N_{\mathcal{T},n},1) & \cdots & \kappa_n(N_{\mathcal{T},n},N_{\mathcal{T},n}) \end{pmatrix},$$

where each $\kappa_n(g,h)$ is a $(\#\mathcal{T}_{g,n})^2\times(\#\mathcal{T}_{h,n})^2$ matrix possibly depending on $\mathcal{W}_n$, and define

$$\hat{\mathbf{\Sigma}}_n(\kappa_n) = \mathrm{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'\kappa_n(g,h)(\hat{\mathbf{u}}_n(h)\otimes\hat{\mathbf{u}}_n(h))\right\}.$$

Also, let $\mathbf{M}_n\circledast_n\mathbf{M}_n$ and $\mathrm{diag}_n(\mathbf{M}_n\circledast_n\mathbf{M}_n)$ be shorthand for

$$\begin{pmatrix} \mathbf{M}_n(1,1)\otimes\mathbf{M}_n(1,1) & \cdots & \mathbf{M}_n(1,N_{\mathcal{T},n})\otimes\mathbf{M}_n(1,N_{\mathcal{T},n}) \\ \vdots & \ddots & \vdots \\ \mathbf{M}_n(N_{\mathcal{T},n},1)\otimes\mathbf{M}_n(N_{\mathcal{T},n},1) & \cdots & \mathbf{M}_n(N_{\mathcal{T},n},N_{\mathcal{T},n})\otimes\mathbf{M}_n(N_{\mathcal{T},n},N_{\mathcal{T},n}) \end{pmatrix}$$

and

$$\begin{pmatrix} \mathbf{M}_n(1,1)\otimes\mathbf{M}_n(1,1) & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{M}_n(N_{\mathcal{T},n},N_{\mathcal{T},n})\otimes\mathbf{M}_n(N_{\mathcal{T},n},N_{\mathcal{T},n}) \end{pmatrix},$$

respectively. Then

$$\begin{aligned}
\hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}} &= \hat{\mathbf{\Sigma}}_n(\mathbf{I}_{N_{\kappa,n}}), \\
\check{\mathbf{\Sigma}}_n^{\mathtt{LZ}} &= \hat{\mathbf{\Sigma}}_n(\mu_n \mathbf{I}_{N_{\kappa,n}}), \\
\hat{\mathbf{\Sigma}}_n^{\mathtt{BR}} &= \hat{\mathbf{\Sigma}}_n(\kappa_n^{\mathtt{BR}}), \qquad \kappa_n^{\mathtt{BR}} = [\mathrm{diag}_n(\mathbf{M}_n^{1/2} \circledast_n \mathbf{M}_n^{1/2})]^{-1}, \\
\hat{\mathbf{\Sigma}}_n^{\mathtt{JK}} &= \hat{\mathbf{\Sigma}}_n(\kappa_n^{\mathtt{JK}}), \qquad \kappa_n^{\mathtt{JK}} = [\mathrm{diag}_n(\mathbf{M}_n \circledast_n \mathbf{M}_n)]^{-1}, \\
\hat{\mathbf{\Sigma}}_n^{\mathtt{CR}} &= \hat{\mathbf{\Sigma}}_n(\kappa_n^{\mathtt{CR}}), \qquad \kappa_n^{\mathtt{CR}} = (\mathbf{M}_n \circledast_n \mathbf{M}_n)^{-1}.
\end{aligned}$$

Remark III.2. An estimator in the spirit of the leave-out estimator of Kline, Saggio, and Sølvsten (2020) is given by

$$\hat{\mathbf{\Sigma}}_n^{\mathtt{LO}} = \mathrm{vec}_d^{-1}\left\{ \frac{1}{n} \sum_{1 \leq g \leq N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(g) \otimes \hat{\mathbf{V}}_n(g))' (\mathbf{y}_n(g) \otimes \mathbf{M}_n(g,g)^{-1} \hat{\mathbf{u}}_n(g)) \right\}.$$

and under specific assumptions, this estimator also serves as a valid estimator. The validity of this estimator and corresponding assumptions will be discussed in the appendix.

III.4 Main Results

Let $\|\cdot\|_\infty$ denote the maximum row sum of its argument.

Theorem III.2. *Suppose Assumptions III.2.1-III.2.2 hold and suppose Assumption III.2.3 holds with*

$$\mathcal{C}_{\mathcal{T},n}^3[\mathcal{C}_{\mathcal{S},n}\rho_n + n(\varrho_n - \rho_n) + n\chi_n\varrho_n] = o(1) \qquad \textit{and} \qquad \frac{\mathcal{C}_{\mathcal{T},n}^4}{n^2} \sum_{1 \leq i \leq n} ||\hat{\mathbf{v}}_{i,n}||^4 = o_p(1).$$

(a) If $\mathcal{C}_{\mathcal{T},n}^3 \mathcal{M}_n = o_p(1)$ *and if*

$$\lim_{\delta \downarrow 0} \liminf_{n \to \infty} \mathbb{P}\left[\min_{1 \leq g \leq N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{-1/2}||_\infty^{-2} > \delta \right] = 1,$$

then (III.3) holds with $\hat{\mathbf{\Sigma}}_n = \hat{\mathbf{\Sigma}}_n^{\mathtt{BR}}$.

(b) If $\mathcal{C}_{\mathcal{T},n}^2 \mathcal{M}_n = o_p(1)$, *then (III.3) holds with* $\hat{\mathbf{\Sigma}}_n \in \left\{ \hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}}, \check{\mathbf{\Sigma}}_n^{\mathtt{LZ}}, \hat{\mathbf{\Sigma}}_n^{\mathtt{JK}} \right\}$.

(c) If

$$\liminf_{\delta \downarrow 0} \liminf_{n \to \infty} \mathbb{P}\left[\min_{1 \leq g \leq N_{\mathcal{T},n}} \left\{ ||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} - \sum_{1 \leq h \leq N_{\mathcal{T},n}, h \neq g} ||\mathbf{M}_n(g,h)||_\infty^2 \right\} > \delta \right] = 1,$$

then (III.3) holds with $\hat{\mathbf{\Sigma}}_n = \hat{\mathbf{\Sigma}}_n^{\mathtt{CR}}$.

(d) If

$$\liminf_{\delta \downarrow 0} \liminf_{n \to \infty} \mathbb{P}\left[\min_{1 \leq g \leq N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} > \delta \right] = 1,$$

then $\hat{\mathbf{\Sigma}}_n^{\mathtt{JK}} \geq \mathbf{\Sigma}_n + o_p(1)$.

If $\mathcal{C}_{\mathcal{T},n} = O(1)$, then the displayed condition of part (a) of Theorem III.2 holds whenever $\mathcal{M}_n = o_p(1)$.

Corollary III.1. *Suppose Assumption III.2.1 holds with* $\mathcal{C}_{\mathcal{S},n} + \mathcal{C}_{\mathcal{T},n} = O(1)$ *and suppose Assumptions III.2.2-III.2.3 hold. If* $\mathcal{M}_n = o_p(1)$, *then (III.3) holds with* $\hat{\mathbf{\Sigma}}_n \in \left\{ \hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}}, \check{\mathbf{\Sigma}}_n^{\mathtt{LZ}}, \hat{\mathbf{\Sigma}}_n^{\mathtt{BR}}, \hat{\mathbf{\Sigma}}_n^{\mathtt{JK}} \right\}$.

Whether or not the cluster sizes are bounded, the displayed conditions of parts (c) and (d) of Theorem III.2 admit sufficient conditions involving only $\mathcal{C}_{\mathcal{T},n}$ and $\mathcal{M}_n$.

Corollary III.2. *Suppose the assumptions of Theorem III.2 are satisfied. If*

$$\liminf_{\delta \downarrow 0} \liminf_{n \to \infty} \mathbb{P}\left[(\mathcal{C}_{\mathcal{T},n}^2 - \mathcal{C}_{\mathcal{T},n} + 2)\mathcal{M}_n + \sqrt{(\mathcal{C}_{\mathcal{T},n} - 1)(1 - \mathcal{M}_n)\mathcal{M}_n} < 1 - \delta \right] = 1,$$

then (III.3) holds with $\hat{\mathbf{\Sigma}}_n = \hat{\mathbf{\Sigma}}_n^{\mathtt{CR}}$.

Corollary III.3. *Suppose the assumptions of Theorem III.2 are satisfied. If*

$$\liminf_{\delta \downarrow 0} \liminf_{n \to \infty} \mathbb{P}\left[\mathcal{C}_{\mathcal{T},n} \mathcal{M}_n < 1 - \delta \right] = 1,$$

then $\hat{\mathbf{\Sigma}}_n^{\mathtt{JK}} \geq \mathbf{\Sigma}_n + o_p(1)$.

Because

$$(\mathcal{C}_{\mathcal{T},n}^2 - \mathcal{C}_{\mathcal{T},n} + 2)\mathcal{M}_n + \sqrt{(\mathcal{C}_{\mathcal{T},n} - 1)(1 - \mathcal{M}_n)\mathcal{M}_n}$$

is an increasing function of $\mathcal{M}_n \in [0, 1/2)$ and $\mathcal{C}_{\mathcal{T},n} \in \mathbb{N}$, there exists a decreasing function $\mathcal{M}^{\mathtt{CR}} : \mathbb{N} \to (0, 1/2]$ such that

$$(\mathcal{C}_{\mathcal{T},n}^2 - \mathcal{C}_{\mathcal{T},n} + 2)\mathcal{M}_n + \sqrt{(\mathcal{C}_{\mathcal{T},n} - 1)(1 - \mathcal{M}_n)\mathcal{M}_n} < 1$$

if and only if

$$\mathcal{M}_n < \mathcal{M}^{\mathtt{CR}}(\mathcal{C}_{\mathcal{T},n}).$$

The function $\mathcal{M}^{\mathtt{CR}}$ satisfies

$$\frac{\sqrt{C - 1 + 4(C^2 - C + 2)} - \sqrt{C - 1}}{2(C^2 - C + 2)} \leq \sqrt{\mathcal{M}^{\mathtt{CR}}(C)} \leq \frac{\sqrt{\frac{C-1}{2} + 4(C^2 - C + 2)} - \sqrt{\frac{C-1}{2}}}{2(C^2 - C + 2)},$$

but does not seem to admit a closed form solution.

Theorem III.2 is silent about the properties of the various variance estimators in the case where the design is cluster-orthogonal in the sense that $\mathbf{M}_n(g, h)$ is a zero matrix whenever $g \neq h$. Indeed, if the design is cluster-orthogonal, then $\mathbf{M}_n(g, g)$ is idempotent for every g, so $\hat{\mathbf{\Sigma}}_n^{\mathtt{BR}}$, $\hat{\mathbf{\Sigma}}_n^{\mathtt{JK}}$, and $\hat{\mathbf{\Sigma}}_n^{\mathtt{CR}}$ are undefined. Moreover, the condition $\mathcal{C}_{\mathcal{T},n}^2 \mathcal{M}_n = o_p(1)$ is violated because

$$\mathcal{C}_{\mathcal{T},n} \mathcal{M}_n \geq 1,$$

the reason being that if $\mathbf{M}_n(g, g) \neq \mathbf{I}_{\#\mathcal{T}_{g,n}}$, then $(\#\mathcal{T}_{g,n})\,\mathcal{M}_n \geq 1$ because

$$\#\mathcal{T}_{g,n} - 1 \geq tr\left[\mathbf{M}_n(g, g)\right] \geq (\#\mathcal{T}_{g,n}) \min_{1 \leq s \leq \#\mathcal{T}_{g,n}} M_{t_{g,n}(s), t_{g,n}(s), n} \geq (\#\mathcal{T}_{g,n})\,(1 - \mathcal{M}_n).$$

In Theorem III.2(a), the purpose of the condition $\mathcal{C}_{\mathcal{T},n}^2 \mathcal{M}_n = o_p(1)$ is to ensure that a key component of the bias of $\hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}}$ is asymptotically negligible. When the design is cluster-orthogonal, the bias component in question is absent and the condition $\mathcal{C}_{\mathcal{T},n}^2 \mathcal{M}_n = o_p(1)$ can therefore be dropped when analyzing $\hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}}$. In the case of

$\hat{\mathbf{\Sigma}}_n^{\mathtt{BR}}, \hat{\mathbf{\Sigma}}_n^{\mathtt{JK}}$, and $\hat{\mathbf{\Sigma}}_n^{\mathtt{CR}}$, arguably the most natural way of accommodating (possibly) singular $\mathbf{M}_n(g,g)$ is to replace matrix inverses with Moore-Penrose inverses. In slight abuse of notation, we therefore define

$$\hat{\mathbf{\Sigma}}_n^{\mathtt{BR}} = \text{vec}_d^{-1}\left\{\frac{1}{n}\sum_{g=1}^{N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'([\mathbf{M}_n(g,g)^{\frac{1}{2}}]^+\hat{\mathbf{u}}_n(g)\otimes[\mathbf{M}_n(g,g)^{\frac{1}{2}}]^+\hat{\mathbf{u}}_n(g))\right\},$$

$$\hat{\mathbf{\Sigma}}_n^{\mathtt{JK}} = \text{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\le g\le N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'(\mathbf{M}_n(g,g)^+\hat{\mathbf{u}}_n(g)\otimes\mathbf{M}_n(g,g)^+\hat{\mathbf{u}}_n(g))\right\},$$

and

$$\hat{\mathbf{\Sigma}}_n^{\mathtt{CR}} = \text{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\le g,h\le N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'\kappa_n^{\mathtt{CR}}(g,h)(\hat{\mathbf{u}}_n(h)\otimes\hat{\mathbf{u}}_n(h))\right\},$$

where

$$\begin{aligned}&\begin{pmatrix}\kappa_n^{\mathtt{CR}}(1,1) & \cdots & \kappa_n^{\mathtt{CR}}(1,N_{\mathcal{T},n})\\ \vdots & \ddots & \vdots\\ \kappa_n^{\mathtt{CR}}(N_{\mathcal{T},n},1) & \cdots & \kappa_n^{\mathtt{CR}}(N_{\mathcal{T},n},N_{\mathcal{T},n})\end{pmatrix}\\ =&\begin{pmatrix}\mathbf{M}_n(1,1)\otimes\mathbf{M}_n(1,1) & \cdots & \mathbf{M}_n(1,N_{\mathcal{T},n})\otimes\mathbf{M}_n(1,N_{\mathcal{T},n})\\ \vdots & \ddots & \vdots\\ \mathbf{M}_n(N_{\mathcal{T},n},1)\otimes\mathbf{M}_n(N_{\mathcal{T},n},1) & \cdots & \mathbf{M}_n(N_{\mathcal{T},n},N_{\mathcal{T},n})\otimes\mathbf{M}_n(N_{\mathcal{T},n},N_{\mathcal{T},n})\end{pmatrix}^+,\end{aligned}$$

and where $(\cdot)^+$ denotes the Moore-Penrose inverse.

With this interpretation, we have $\hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}} = \hat{\mathbf{\Sigma}}_n^{\mathtt{BR}} = \hat{\mathbf{\Sigma}}_n^{\mathtt{JK}} = \hat{\mathbf{\Sigma}}_n^{\mathtt{CR}}$ when the design is cluster-orthogonal.

Theorem III.3. *Suppose Assumptions III.2.1-III.2.2 hold and suppose Assumption III.2.3 holds with*

$$\mathcal{C}_{\mathcal{T},n}^3[\mathcal{C}_{\mathcal{S},n}\rho_n + n(\varrho_n-\rho_n) + n\chi_n\varrho_n] = o(1) \qquad \textit{and} \qquad \frac{\mathcal{C}_{\mathcal{T},n}^4}{n^2}\sum_{1\le i\le n}||\hat{\mathbf{v}}_{i,n}||^4 = o_p(1).$$

If the design is cluster-orthogonal, then (III.3) holds with $\hat{\mathbf{\Sigma}}_n \in \left\{\hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}}, \hat{\mathbf{\Sigma}}_n^{\mathtt{BR}}, \hat{\mathbf{\Sigma}}_n^{\mathtt{JK}}, \hat{\mathbf{\Sigma}}_n^{\mathtt{CR}}\right\}$. *If also* $\mathcal{C}_{\mathcal{T},n}\mathcal{M}_n = o_p(1)$, *then (III.3) holds with* $\hat{\mathbf{\Sigma}}_n = \check{\mathbf{\Sigma}}_n^{\mathtt{LZ}}$.

III.5 Simulation

III.5.1 Simulation Design

We conduct a simulation study to assess the finite sample properties of the standard errors we proposed here and compared them with other cluster robust standard errors available in the literature. Based on the regression model III.1, we consider a linear data generating process with growing dimensions that includes a series of different standard errors including the well known Liang-Zeger estimator for clusters (LZ) with and without degree of freedom adjustment (LZ-df) that degenerates to HC0 and HC1 standard error when each cluster has only one unit. We also include the 'bias-reduction' standard error (BR) and the 'jackknife' standard error (JK) that are clustered versions of HC2 and HC3 estimators. We also contain the cluster robust standard error we proposed here (CR), and the leave-out standard error (LO).

Our paper presents theory for Gaussian-based inference methods and for each inference method we report both empirical coverage and their average interval length. The latter provides a summary of efficiency for each inference method.

The Gaussian-based confidence interval takes the form

$$I_l = \left[\hat{\beta} - \Phi^{-1}(1-\alpha/2)\sqrt{\frac{\hat{\Omega}_l}{n}}, \hat{\beta} - \Phi^{-1}(\alpha/2)\sqrt{\frac{\hat{\Omega}_l}{n}}\right] \qquad \hat{\Omega}_l = \hat{\Gamma}^{-1}\hat{\Sigma}_l\hat{\Gamma}^{-1}$$

where Φ^{-1} denotes the inverse of the c.d.f of the Gaussian distribution and $\hat{\Sigma}_l$ with $l \in$ {LZ,LZ-df,BR,JK,LO,CR} corresponds to each of the variance estimator discussed in the paper.

The data generating process for the linear regression model with many covariates follows Cattaneo et al. (2018). For simplicity, we drop the subindex n in the

discussion below and the data follow

$$\begin{aligned} y_i &= \beta x_i + \boldsymbol{\gamma_n'}\mathbf{w}_i + u_i, \qquad i = 1, \ldots, n, \\ x_i|\mathbf{w}_i &\sim \mathcal{N}(0, \sigma_{x,i}^2) \qquad \sigma_{x,i}^2 = \varkappa_x(1 + (\boldsymbol{\iota}'\mathbf{w}_i)^2) \end{aligned}$$

The data generating process of the error term $u_{i,n}$ follows MacKinnon, Nielsen, Webb, et al. (2020). For each cluster of the error term, we define

$$\mathbf{u}_n(g) = (u_{t_{g,n}(1)}, \ldots, u_{t_{g,n}(\#\mathcal{T}_{g,n})})', \qquad g = 1, 2, \ldots, N_{\mathcal{T},n}$$

which follows

$$\mathbf{u}_n(g) = \mathbf{P}_\xi \boldsymbol{\xi}_n(g) + p_\varepsilon \boldsymbol{\varepsilon}_n(g), \quad \boldsymbol{\varepsilon}_n(g) \sim \mathcal{N}\left(\mathbf{0}, \text{diag}\{\sigma^2_{u,t_{g,n}(1)}, \ldots, \sigma^2_{u,t_{g,n}(\#\mathcal{T}_{g,n})}\}\right),$$

$\boldsymbol{\xi}_n(g) = [\xi_{g1,n}, \ldots, \xi_{gJ,n}]$ is a J-vector of unobserved random factors. The $\#\mathcal{T}_{g,n} \times J$ loading matrix $\mathbf{P}_\xi$ has (i,j)-th entry $p_\xi \mathbb{I}\left(j = \lfloor (i-1)J/\#\mathcal{T}_{g,n} \rfloor + 1\right)$ where $\lfloor . \rfloor$ denotes the integer part of the argument. When J=1, entries of the loading matrix $\mathbf{P}_\xi$ are ones and the error term degenerates to the commonly known random effect model. In order to avoid the situation where cluster fixed effect fully captured the correlation within each cluster, we require that the J unobserved random factors are also correlated with each other following

$$\xi_{g1,n} \sim \mathcal{N}(0,1), \qquad \xi_{gj,n} = \rho \xi_{gj-1,n} + e_{gj,n} \qquad e_{gj,n} \sim \mathcal{N}(0, 1-\rho^2), \qquad j = 2, \ldots, J$$

$\boldsymbol{\varepsilon}_n(g)$ is a noise term with independent normal distribution and conditional heterogeneous variance where

$$\sigma_{u,i}^2 = \varkappa_u (t(x_i) + \boldsymbol{\iota}'\mathbf{w}_i)^2$$

Regarding the selection of parameters, we have that $\boldsymbol{\iota} = (1, 1, \ldots, 1)'$, $d = 1$, $\beta = 1$, $\boldsymbol{\gamma_n} = \mathbf{0}$, $t(a) = a\mathbb{I}(-2 \leq a \leq 2) + 2\text{sgn}(a)(1 - \mathbb{I}(-2 \leq a \leq 2))$ and $\varkappa_x$ and $\varkappa_u$ are constant that makes $\mathbb{V}[x_{t_{G,n(1)}}] = \mathbb{V}[u_{t_{g,n(1)}}] = 1$. We also pick $\rho = 0.5$, $p_\xi = 0.7$, $p_\varepsilon = \sqrt{1 - p_\xi^2}$ and J=3 for error terms.

For the generation of covariates, we also allow them to have certain form of correlation and we allow the cluster design for covariates to be different from that of the error term. If the regressors are independent, then we build them based on $\mathbf{w_i} \overset{\text{i.i.d}}{\sim} \mathbf{U}[-1,1]^K$. If we would like to impose some correlation structure on the covariates, the data generating process of regressors is almost the same as that of the error term where for the k-th regressor

$$\mathbf{w}_n^k(G) = (w^k_{t_{G,n}(1)}, \ldots, w^k_{t_{G,n}(\#\mathcal{S}_{G,n})})', \qquad G = 1, 2, \ldots, N_{\mathcal{S},n}$$

has the form

$$\mathbf{w}_n^k(G) = \mathbf{L}_\lambda \boldsymbol{\lambda}_n^k(G) + l_\epsilon \boldsymbol{\epsilon}_n^k(G), \qquad \boldsymbol{\epsilon}_i^k \overset{\text{i.i.d}}{\sim} \mathbf{U}[-1,1], \qquad G = 1, 2, \ldots, N_{\mathcal{T},n}$$

$\boldsymbol{\lambda}_n^k(g) = [\lambda^k_{g1,n}, \ldots, \lambda^k_{gJ,n}]$ is a J-vector of unobserved random factors. The $\#\mathcal{S}_{G,n} \times J$ loading matrix $\mathbf{L}_\lambda$ has (i,j)-th entry $l_\lambda \mathbb{I}\left(j = \lfloor (i-1)J/\#\mathcal{S}_{G,n} \rfloor + 1\right)$. The J unobserved random factors are also correlated with each other following

$$\lambda_{G1,n} \sim \mathbf{U}[-1,1], \quad \lambda_{Gj,n} = \rho\lambda_{Gj-1,n} + \tilde{e}_{Gj,n} \quad \tilde{e}_{Gj,n} \sim \sqrt{1-\rho^2}\mathbf{U}[-1,1], \quad j = 2, \ldots, J$$

We pick J=10 for covariates, and choose $l_\lambda = 0.7$, $l_\epsilon = \sqrt{1 - l_\lambda^2}$.

In order to analyze the effect of the cluster design on the performance of different standard errors, we conduct simulation designs on both homogeneous and heterogeneous error term clusters. Under homogeneous design, all the clusters have the same size with $\#\mathcal{T}_{g,n} = 6$ while for the heterogeneous case half of the clusters have a cluster size of 4 and half of them have a cluster size of 8. For the regressors, we allow them either to be independent or follow the correlated data generation process mentioned above with $\#\mathcal{S}_{G,n} = 24$. We study $s = 1000$ simulations to study the finite sample performance of different variance estimators on a sample whose size is $n = 600$. In total there should be 4 models with different choices of homogeneous or heterogeneous cluster sizes and independent or dependent regressors and for each of the models, we allow the number of regressors to grow at the same rate as the sample size. We consider five dimensions of K with $\{1, 1 + 0.1n, \ldots, 1 + 0.4n\}$ and

the results are shown below.

III.5.2 Results and Discussions

In practice, one thing worth mentioning is that as the estimator is not guaranteed to be nonnegative by construction like the HC-k estimator, we need to make some finite sample adjustments to avoid calculating the square root of a negative term. This will happen to the variance estimator calculated from each cluster as well as the total variance. In our setup, we try three different regularization for Cluster Robust estimator(CR) and Leave-One-Out(LO) estimators and list the result separately.

$$\hat{\Sigma}_n^{LO} = \frac{1}{n}\sum_{g=1}^{\mathcal{N}_{\mathcal{T},n}} \hat{\mathbf{V}}_n(g)'\mathbf{y}_n(g)\mathbf{M}_n(g,g)^{-1}\hat{\mathbf{u}}(g)\hat{\mathbf{V}}_n(g)$$

$$\hat{\Sigma}_n^{CR} = \frac{1}{n}\sum_{g=1}^{\mathcal{N}_{\mathcal{T},n}} \hat{\mathbf{V}}_n(g)' \sum_{h=1}^{\mathcal{N}_{\mathcal{T},n}} \boldsymbol{\kappa}_n(g,h)\hat{\mathbf{u}}_n(h)\hat{\mathbf{u}}_n(h)'\hat{\mathbf{V}}_n(g)$$

Adjustment 1 Drop Negative $\hat{\Sigma}_n$ and only calculate $\mathbb{P}[coverage|\hat{\Sigma}_n > 0]$.
Adjustment 2 For those have negative $\hat{\Sigma}_n$, we calculate

$$\tilde{\Sigma}_n^{LO}(g) = \max\{\hat{\mathbf{V}}_n(g)'\mathbf{y}_n(g)\mathbf{M}_n(g,g)^{-1}\hat{\mathbf{u}}(g)\hat{\mathbf{V}}_n(g), 0\}$$

$$\tilde{\Sigma}_n^{CR}(g) = \max\{\hat{\mathbf{V}}_n(g)' \sum_{h=1}^{\mathcal{N}_{\mathcal{T},n}} \boldsymbol{\kappa}_n(g,h)\hat{\mathbf{u}}_n(h)\hat{\mathbf{u}}_n(h)'\hat{\mathbf{V}}_n(g), 0\}$$

for each cluster and then calculate $\hat{\Sigma}_n = \sum_{g=1}^{\mathcal{N}_{\mathcal{T},n}} \tilde{\Sigma}_n(g)$
Adjustment 3 Implement Adjustment 2 not only for $\hat{\Sigma}_n < 0$ but for all $\hat{\Sigma}_n$.

We also list two negative ratios, the first negative ratio measure $\mathbb{P}[\hat{\Sigma}_n < 0]$ and the second ratio measures $\mathbb{P}[\tilde{\Sigma}_n(g) = 0]$ among all clusters for reference.

Simulation results for four different models with 1000 replications are summarized below. For each of them, I include the performance of LZ-estimator, LZ-df, BR, JK estimator, the performance of leave-one-out estimator and cluster robust estimator we propose with three different kinds of finite adjustment. We also list two different

kinds of negative ratio for leave-one-out and cluster-robust estimator.

Table III.1: Simulation Results, Independent Regressors, n=600, S=1000
Homogeneous Cluster Size

	(a): Empirical Coverage									
	LZ	HC1	HC2	HC3	LO1	LO2	LO3	CR1	CR2	CR3
K/n=0.002	0.956	0.958	0.956	0.957	0.952	0.950	0.986	0.956	0.956	0.956
K/n=0.102	0.922	0.937	0.937	0.947	0.941	0.940	0.971	0.939	0.939	0.942
K/n=0.202	0.886	0.924	0.922	0.950	0.910	0.913	0.968	0.931	0.931	0.934
K/n=0.302	0.855	0.918	0.919	0.972	0.932	0.932	0.975	0.934	0.934	0.950
K/n=0.402	0.824	0.911	0.911	0.974	0.922	0.922	0.966	0.934	0.934	0.954
	(b): Interval Length									
	LZ	HC1	HC2	HC3	LO1	LO2	LO3	CR1	CR2	CR3
K/n=0.002	0.192	0.193	0.192	0.192	0.191	0.191	0.229	0.192	0.192	0.192
K/n=0.102	0.264	0.280	0.279	0.295	0.287	0.286	0.333	0.286	0.286	0.288
K/n=0.202	0.245	0.275	0.274	0.307	0.285	0.285	0.336	0.287	0.287	0.291
K/n=0.302	0.237	0.284	0.283	0.338	0.305	0.304	0.357	0.302	0.302	0.312
K/n=0.402	0.241	0.313	0.311	0.402	0.341	0.340	0.396	0.337	0.337	0.358
	(c): Negative Ratio									
	LONeg1			CRNeg1			LONeg2			CRNeg2
K/n=0.002	0.000			0.000			0.465			0.037
K/n=0.102	0.001			0.000			0.394			0.239
K/n=0.202	0.005			0.000			0.386			0.325
K/n=0.302	0.001			0.000			0.380			0.397
K/n=0.402	0.001			0.000			0.368			0.454

From the tables we can see that as the dimension of covariates increasing, the performance of commonly used LZ estimator and LZ-df, BR standard errors perform worse and have a relatively low empirical coverage. JK standard error is always conservative. For the two standard errors we propose in our paper, Cluster-robust standard error is always nonnegative if we take the sum while the LO standard error will be negative but with a relatively low ratio. If we consider the estimator for each cluster separately, the negative ratio is huge. What is surprising is the third adjustment for CR estimator, performs really well although it always drives LO estimator too conservative. The CR estimator exhibits a consistent improvement against other estimators while the performance of LO standard error is not quite ideal. The overall negative ratio for the standard error is not huge while the negative ratio for each cluster is relatively large but on average the performance is fine. The

Table III.2: Simulation Results, Independent Regressors, n=600, S=1000
Heterogeneous Cluster Size

	(a): Empirical Coverage									
	LZ	LZ-df	BR	JK	LO1	LO2	LO3	CR1	CR2	CR3
K/n=0.002	0.946	0.946	0.946	0.946	0.945	0.942	0.984	0.946	0.946	0.946
K/n=0.102	0.924	0.941	0.941	0.956	0.921	0.923	0.974	0.948	0.948	0.950
K/n=0.202	0.881	0.928	0.926	0.950	0.933	0.934	0.962	0.934	0.934	0.939
K/n=0.302	0.861	0.930	0.930	0.967	0.923	0.921	0.972	0.940	0.940	0.950
K/n=0.402	0.823	0.907	0.903	0.968	0.917	0.920	0.977	0.931	0.931	0.948
	(b): Interval Length									
	LZ	LZ-df	BR	JK	LO1	LO2	LO3	CR1	CR2	CR3
K/n=0.002	0.192	0.193	0.192	0.193	0.192	0.192	0.232	0.192	0.192	0.192
K/n=0.102	0.265	0.280	0.279	0.295	0.284	0.284	0.333	0.287	0.287	0.288
K/n=0.202	0.265	0.298	0.297	0.332	0.312	0.312	0.361	0.311	0.311	0.316
K/n=0.302	0.242	0.291	0.289	0.346	0.308	0.307	0.363	0.309	0.309	0.319
K/n=0.402	0.206	0.268	0.266	0.344	0.290	0.291	0.354	0.288	0.288	0.306
	(c): Negative Ratio									
	LONeg1			CRNeg1			LONeg2			CRNeg2
K/n=0.002	0.000			0.000			0.460			0.039
K/n=0.102	0.005			0.000			0.387			0.239
K/n=0.202	0.004			0.000			0.367			0.329
K/n=0.302	0.001			0.000			0.371			0.400
K/n=0.402	0.006			0.000			0.385			0.455

performance of different standard errors are quite consistent across either change in cluster size or change in dependent structures of covariates.

III.6 Conclusion

In this paper we established asymptotic normality results of the OLS-based estimator when we possibly have many covariates in the sense of non-ignorable compared with the sample size and have correlations between error terms with a clustering structure. Starting from there, we investigate the performance of a series of cluster-robust standard errors under this high dimensional setup. We analyze and provide different conditions for previously used cluster robust standard errors to be effective and point out a trade-off between the cluster designs, mainly on the largest cluster size, and the dimension of the covariates for the purpose of keeping the validity. We also propose a new formula for the standard error that is robust to the existence of

Table III.3: Simulation Results, Dependent Regressors, n=600, S=1000
Homogeneous Cluster Size

	(a): Empirical Coverage									
	LZ	LZ-df	BR	JK	LO1	LO2	LO3	CR1	CR2	CR3
K/n=0.002	0.964	0.965	0.965	0.965	0.966	0.962	0.990	0.965	0.965	0.965
K/n=0.102	0.923	0.937	0.939	0.956	0.913	0.912	0.974	0.946	0.946	0.948
K/n=0.202	0.876	0.919	0.918	0.946	0.905	0.906	0.958	0.927	0.927	0.932
K/n=0.302	0.854	0.928	0.927	0.965	0.917	0.919	0.972	0.938	0.938	0.947
K/n=0.402	0.825	0.917	0.909	0.971	0.918	0.920	0.969	0.924	0.924	0.946
	(b): Interval Length									
	LZ	LZ-df	BR	JK	LO1	LO2	LO3	CR1	CR2	CR3
K/n=0.002	0.189	0.190	0.190	0.190	0.190	0.189	0.227	0.190	0.190	0.190
K/n=0.102	0.253	0.268	0.268	0.284	0.271	0.271	0.324	0.274	0.274	0.276
K/n=0.202	0.226	0.254	0.253	0.285	0.262	0.262	0.320	0.265	0.265	0.269
K/n=0.302	0.242	0.291	0.287	0.347	0.305	0.305	0.362	0.306	0.306	0.318
K/n=0.402	0.229	0.297	0.291	0.378	0.316	0.316	0.378	0.313	0.313	0.335
	(c): Negative Ratio									
	LONeg1			CRNeg1			LONeg2			CRNeg2
K/n=0.002	0.000			0.000			0.466			0.038
K/n=0.102	0.004			0.000			0.400			0.259
K/n=0.202	0.007			0.000			0.397			0.347
K/n=0.302	0.006			0.000			0.380			0.414
K/n=0.402	0.006			0.000			0.382			0.465

clustering and high-dimensional covariates under fewer restrictions. Sufficient conditions to make this standard error work are also provided and numerical evidence is also presented.

Although closely connected, our results are not trivially generalized from linear models with many covariates with independent errors. Following the idea of keeping the cluster structure flexible, we do not impose restrictions on how units are correlated with each other within clusters. What is more, we do not restrict the cluster size to be finite and allow that the covariates and error terms have different correlation structures. The trade-off between the largest cluster size and the number of covariates brings more flexibilities and also makes the challenge caused by many covariates happen more easily when we have relatively large clusters.

Further, although this approach is conducted under the structure of one-way clustering, I would expect this method can be generalized to multiway-clustering setup to incorporate more empirically related setup. It will also be interesting to

Table III.4: Simulation Results, Dependent Regressors, n=600, S=1000
Heterogeneous Cluster Size

	(a): Empirical Coverage									
	LZ	LZ-df	BR	JK	LO1	LO2	LO3	CR1	CR2	CR3
K/n=0.002	0.946	0.948	0.947	0.947	0.950	0.944	0.977	0.946	0.946	0.946
K/n=0.102	0.919	0.933	0.933	0.948	0.913	0.914	0.972	0.938	0.938	0.940
K/n=0.202	0.872	0.910	0.908	0.939	0.900	0.899	0.962	0.921	0.921	0.926
K/n=0.302	0.852	0.921	0.918	0.967	0.907	0.910	0.978	0.932	0.932	0.941
K/n=0.402	0.817	0.913	0.901	0.960	0.910	0.910	0.966	0.915	0.915	0.943
	(b): Interval Length									
	LZ	LZ-df	BR	JK	LO1	LO2	LO3	CR1	CR2	CR3
K/n=0.002	0.188	0.189	0.189	0.189	0.189	0.188	0.229	0.189	0.189	0.189
K/n=0.102	0.250	0.265	0.265	0.281	0.268	0.268	0.323	0.271	0.271	0.273
K/n=0.202	0.250	0.280	0.279	0.315	0.291	0.290	0.347	0.292	0.292	0.297
K/n=0.302	0.225	0.270	0.267	0.322	0.283	0.284	0.345	0.284	0.284	0.295
K/n=0.402	0.227	0.294	0.287	0.373	0.310	0.310	0.375	0.309	0.309	0.331
	(c): Negative Ratio									
	LONeg1			CRNeg1			LONeg2			CRNeg2
K/n=0.002	0.000			0.000			0.458			0.039
K/n=0.102	0.009			0.000			0.393			0.264
K/n=0.202	0.006			0.000			0.379			0.355
K/n=0.302	0.009			0.000			0.383			0.418
K/n=0.402	0.007			0.000			0.379			0.467

think about the inclusion of many covariates in other related setup, for example, the time series models.

CHAPTER IV

Robust Pricing Under Strategic Trading

IV.1 Introduction

Economic theory often assumes that the joint distribution of random variables is common knowledge. For example, participants in financial markets know the distribution of asset returns, employers know the distribution of employees' unobserved ability, and consumers know the distribution of a new product's quality. In practice, however, such distributions are often unknown, in which case information will be difficult to process even if it is public and precise. For instance, when a publicly traded company discloses information or a central bank cuts interest rates, it is often unclear to many market participants how much of the information is already anticipated and priced in.

When the distribution is unknown, people face *ambiguity* (see Ellsberg (1961)). A seminal paper by Gilboa and Schmeidler (1989) characterizes an axiomatic model that exhibits ambiguity aversion, the *maxmin expected utility* model. In this model, people maximize the *worst-case* (across all plausible distributions) payoff guarantee; that is, people behave in an optimal way that is robust to the unknown distribution. This model and, more generally, the maxmin principle have been widely used in many applications (see Section 8).

This chapter is based on the paper "Robust Pricing Under Strategic Trading"(Gong, Ke, Qiu, and Shen (2022))

We follow this approach and examine how people react to *public* information in a simple model of strategic trading. Specifically, an asset is traded in the market. Market participants are the *probabilistically informed trader*, the market maker, and liquidity traders. None of them has private information about the value of the asset v. They receive the same signal s from a public event at the beginning. Then, on each of the trading dates, all traders submit market orders to trade and the market maker determines a price of the asset at which the orders are traded. The probabilistically informed trader trades to maximize profit, and liquidity traders trade for idiosyncratic reasons.

After the public event, the probabilistically informed trader (she) can update her belief about v based on the joint distribution of v and s. The joint distribution is normal and the mean of s is $\bar{s}$.[1] However, not all market participants know how to translate s into information about v. As noted above, it is often unclear what public information should have been anticipated, or even whether the information is positive or negative given the price. Therefore, we consider a situation in which the market maker (he) knows everything about the joint distribution except $\bar{s}$—that is, what to expect from s in the first place.

Without knowing $\bar{s}$, the market maker does not know how to form the posterior of v based on the public information, but he prices the asset in a way that is robust to the unknown $\bar{s}$—he chooses a pricing strategy that has the best worst-case (across all possible values of $\bar{s}$) payoff. We assume that the market maker wants to set a fair price for traders, and his payoff is the sum of his current and future price errors (differences between prices and v) evaluated according to some general loss function. Due to the asymmetric information about $\bar{s}$, the probabilistically informed trader may manipulate her orders to affect the market maker's behavior, and the market maker's aversion to mispricing determines how she takes the orders into account as she sets prices.

We study the linear equilibrium of such a model, called the *dynamic linear robust*

[1]The normal distribution can be replaced with the more general elliptical distribution, which could allow for thick-tailed distributions. We discuss this after Proposition IV.2.

pricing (RP) equilibrium.[2] First, we show that given any linear trading strategy of the probabilistically informed trader, the market maker's *backward-induction* optimal robust linear pricing strategy is equivalent to the following two-step learning procedure.[3] On each trading date, based on the received orders (and the initial public signal), the market maker first estimates the unknown parameter $\bar{s}$ optimally using the *best linear unbiased estimator* (BLUE), which depends on the probabilistically informed trader's strategy. Then, the market maker uses the estimated joint distribution to update his belief about v and lets the price be the conditional expectation of v. The two-step learning procedure resembles how people deal with unknown distributions in practice: They often estimate the distribution. As Hansen (2007) emphasizes, it is important to understand how real-time distribution estimation affects people's behavior and equilibrium outcomes.

This result enables us to characterize the unique dynamic linear RP equilibrium indirectly by characterizing the unique *dynamic BLUE* equilibrium. In a dynamic BLUE equilibrium, instead of finding the optimal robust linear pricing strategy, the market maker is assumed to adopt the two-step learning procedure. Every dynamic linear RP equilibrium is equivalent to a dynamic BLUE equilibrium, and vice versa.

The indirect characterization of the dynamic linear RP equilibrium is useful in many ways. First, a key variable in the dynamic BLUE equilibrium is the BLUE of $\bar{s}$, which does not appear in the definition of the dynamic linear RP equilibrium. The BLUE of $\bar{s}$ can be thought of as an auxiliary variable in the dynamic linear RP equilibrium, which turns out to help us understand the structure of the equilibrium better—not only the market maker's behavior, but also the probabilistically informed trader's.

Second, we show that the dynamic linear RP equilibrium exhibits two properties by analyzing the dynamic BLUE equilibrium. First, on average, the equilibrium prices exhibit *underreaction* under the true joint distribution; that is, the expected

[2]The linearity assumption is common in models based on Kyle (1985). This assumption might be descriptively appealing, because linear strategies are parsimonious for the probabilistically informed trader and the market maker to use. Technically, it makes our analysis tractable.

[3]Mathematically, this characterization extends some results from the literature on minimax statistical estimation to a dynamic setting using a different proof strategy.

prices are less sensitive to public information than in a benchmark model in which the market maker knows $\bar{s}$, and as time goes by the expected prices move toward the price in the benchmark model. Classic economic theories posit that the market is efficient and public information will rapidly be fully reflected in prices. The benchmark model is consistent with this. In sharp contrast, a large amount of empirical evidence finds underreaction to public events. For example, stock returns often experience post-earnings announcement drift.[4] Thus, our theory provides an explanation for underreaction.[5]

The second property of the dynamic linear RP equilibrium addresses a basic question. Without knowing $\bar{s}$, the market maker faces ambiguity. What does the market maker do with ambiguity as he solves the dynamic robust pricing problem? Will ambiguity be eliminated? We show that if the trading frequency is arbitrarily high, the market maker will learn $\bar{s}$ in the dynamic BLUE equilibrium in the end—or, equivalently, as the market maker implements the optimal robust pricing strategy in the dynamic linear RP equilibrium, eventually $\bar{s}$ will be revealed in the price and the public information will be fully incorporated into the price.

Last, we examine how the above findings rely on our assumptions about the market maker's behavior. First, we compare our model to a model with a Bayesian market maker. The Bayesian market maker also does not know $\bar{s}$ but treats it as a random variable and has a prior over it. We show that underreaction does not always arise in the Bayesian case. Second, we analyze how the equilibrium behavior changes if the market maker uses some estimator of $\bar{s}$ other than the BLUE in the first step of the two-step learning procedure. We find that multiple equilibria exist, and the probabilistically informed trader does not necessarily benefit from the fact

[4]Ball and Brown (1968) and Beaver (1968) first document the post-earnings announcement drift, whose robustness is confirmed in many studies (see Bernard and Thomas (1989, 1990); Bernard (1992); Chan, Jegadeesh, and Lakonishok (1996); and Hou, Chen, and Zhang (2018)). For underreaction to other public events, see Ikenberry, Lakonishok, and Vermaelen (1995); Michaely, Thaler, and Womack (1995); Hong, Lim, and Stein (2000); Hilary and Shen (2013); and Ng, Tuna, and Verdi (2013).

[5]Several theories featuring limited attention, momentum traders, short-selling constraints, or other financial constraints have been proposed to explain underreaction. See Daniel, Hirshleifer, and Subrahmanyam (1998); Hong and Stein (1999); Frazzini (2006); and Hirshleifer, Lim, and Teoh (2009), among others.

that the market maker uses a "suboptimal" estimator.

The structure of the paper is as follows. Section 2 describes the setup that applies to all models in our paper. The benchmark model is introduced in Section 3, the static model in Section 4, and the dynamic model in Section 5. Section 6 studies underreaction and market efficiency in the dynamic model. In Section 7, we analyze how our findings depend on the assumptions about the market maker's behavior. Section 8 discusses related literature, and Section 9 concludes.

IV.2 The Setup

An asset is traded in the market. Market participants consist of a *probabilistically informed trader*, a market maker, and liquidity traders.[6] The value of the asset v, a normally distributed random variable with mean $\bar{v}$ and variance σ_v^2, will be revealed sometime after the end of all trades. The distribution of the value of the asset is common knowledge to all market participants, and none of the participants will ever have private information about v.

Before any trading begins there is a public event, from which all participants receive the same public signal s. The signal is informative about the value of the asset only if a market participant knows the joint distribution of v and s. Upon receiving the signal, the probabilistically informed trader (she) updates her belief according to the joint distribution of v and s, $\mathcal{N}\left(\begin{bmatrix} \bar{v} \\ \bar{s} \end{bmatrix}, \begin{bmatrix} \sigma_v^2 & \rho\sigma_v\sigma_s \\ \rho\sigma_v\sigma_s & \sigma_s^2 \end{bmatrix}\right)$.[7] We assume $\rho > 0$ throughout the paper, since the other case is symmetric. The market maker (he) knows that the probabilistically informed trader knows the joint distribution, but he may or may not know the joint distribution. In the benchmark model, he knows. In our main model, he knows that the joint distribution is normal and all parameters of the distribution *except for* $\bar{s}$.[8]

[6] Sometimes the number of liquidity traders is important (see, for example, Han, Tang, and Yang (2016)), but in our model it is not.

[7] We discuss how the normal distribution assumption can be relaxed after Proposition IV.2.

[8] The assumption that the market maker knows the covariance matrix but not $\bar{s}$ may be understood as follows. Suppose the public information is disclosed from a source (e.g., a firm) that the market maker is familiar with. Then, he may have a good understanding of how noisy the signal

Following the public event, trading begins. Our main model, called the dynamic model, has multiple trading dates. On each trading date, the following sequence of events takes place. First, based on the (only) public signal and past prices, the probabilistically informed trader submits a market order at the same time as liquidity traders submit their orders. A market order specifies the quantity of the asset a trader commits to trade at a price that will be determined by the market maker. The probabilistically informed trader is risk-neutral and trades to maximize her profit. Liquidity traders trade for idiosyncratic reasons, and their (total) order is a normally distributed random variable that is independent of all other random variables. These are common knowledge.

Next, the market maker observes the total order, which consists of the probabilistically informed trader's order and liquidity traders' orders. He trades the quantity that clears the market at a price determined by him according to some criterion that will be elaborated on later. Note that the market maker cannot differentiate between the probabilistically informed trader's order and liquidity traders' orders. Nonetheless, the total order is potentially useful for him to set the optimal price. Again, these are common knowledge.

The above setup applies to all models in our paper. The benchmark model mentioned above will be introduced in Section 3, which will help us understand market participants' behavior when $\bar{s}$ is known to everyone and will help us define underreaction later. Then, before introducing the dynamic model in which the market maker does not know $\bar{s}$ in Section 5, we will first present in Section 4 its static special case. The static model will help us explain some key ingredients of the dynamic model in a simple setting. Finally, the models in Section 7, which are variations or extensions of the static model, will also use the same setup, except that we will make different assumptions about what the maket maker knows or does.

---

typically is and how it usually correlates with the fundamental. The particular event underlying the signal s, however, is new. We may assume that the variance and covariance of this signal are the same as before and therefore the market maker knows them, but he will not know what $\bar{s}$ "usually" is for this new event.

IV.3 The Benchmark

Our benchmark model assumes that the market maker knows the joint distribution of v and s. At $t = 0$, every market participant observes s. For a reason that will soon become clear, we assume without loss of generality that there is only one trading date at $t = 1$. For any s, the probabilistically informed trader's order is $X(s)$. The liquidity traders' order is u, a normally distributed random variable with mean 0 and variance σ_u^2 that is independent of all other random variables. Let $y = X(s)+u$ denote the total order. Based on s and y, the market maker determines the price at which the orders are traded, $P(s, y)$. When v is revealed after $t = 1$, the probabilistically informed trader receives profit $\pi = (v - P(s, y))X(s)$.

The following equilibrium notion is from Kyle (1985), except that our model does not have a trader who knows v but has a public signal s.

Definition IV.1. *The pair of functions X and P is an equilibrium if*

1. *given $X(s)$, $P(s, y) = \mathbb{E}[v|s, y]$; and*
2. *given $P(s, y)$, $X(s)$ maximizes $\mathbb{E}[\pi|s]$.*

The probabilistically informed trader is risk-neutral and maximizes the expected profit.[9] The market maker sets the price equal to the conditional expectation of v. Note that assuming $P(s, y) = \mathbb{E}[v|s, y]$ is equivalent to assuming that the market maker's goal is to find a pricing strategy that minimizes the *mean squared price error function*; that is, $P(s, y)$ solves

$$\min_{\tilde{P}(s,y)} \mathbb{E}[(\tilde{P}(s, y) - v)^2]. \tag{IV.1}$$

In other words, the market maker wants the price to deviate from v as little as possible, measured by the mean squared price error function.[10] Therefore, Definition

[9] See Subrahmanyam (1991) and Holden and Subrahmanyam (1994) for models that relax the risk neutrality assumption.

[10] See footnote 12 for discussion of an alternative way to interpret the market maker's behavior.

IV.1 is equivalent to the following definition, and this equivalence will be useful in the next section.

Definition IV.2. *The pair of functions X and P is an equilibrium if*

1. *given $X(s)$, $P(s,y)$ solves (IV.1); and*
2. *given $P(s,y)$, $X(s)$ maximizes $\mathbb{E}[\pi|s]$.*

The total order y does not contain any information beyond s, and the market maker already knows s. Therefore, Definition IV.1 implies that

$$P(s,y) = \mathbb{E}[v|s,y] = \mathbb{E}[v|s] = \bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}).$$

Given this,

$$\mathbb{E}[\pi|s] = \mathbb{E}[(v - P(s,y))X(s)|s] = 0$$

for any $X(s)$. Hence, in any equilibrium, following the public event, the price should be immediately adjusted to $\mathbb{E}[v|s]$, which fully reflects the information from the public signal s.[11] We summarize these observations below and omit the proof.

Proposition IV.1. *The pair of functions X and P is an equilibrium if and only if $P(s,y) = \mathbb{E}[v|s]$.*

It is straightforward to verify that even if there are multiple trading dates, the price will become $\mathbb{E}[v|s]$ on the first trading date and remain unchanged. Moreover, $P(s,y) = \mathbb{E}[v|s]$ even if the market maker receives a zero total order; that is, the benchmark model does not require any order for the price adjustment to be completed.

IV.4 A Special Case: The Static Model

From here on, we assume that the market maker does not know $\bar{s}$ and focus on linear strategies as in Kyle (1985); that is, the trading strategy and the pricing

[11] The price of the asset before the public event is $\bar{v}$.

strategy are both affine functions. Before introducing the dynamic model, we examine a special case in which the only trading date is at $t = 1$. The probabilistically informed trader's order is given by her linear trading strategy $X(\bar{s}, s)$. To emphasize that she is the only market participant who knows $\bar{s}$, we write X as a function of both s and $\bar{s}$. The assumption on liquidity traders' order u is the same as before and $y = X(\bar{s}, s) + u$ is the total order.

To model the market maker's behavior, we follow the approach from the literature on ambiguity and robust contract and mechanism design, which often involves people confronting unknown distributions. The main assumption is that the strategy that has the best *worst-case* payoff guarantee will be adopted in this situation. This model is axiomatized by Gilboa and Schmeidler (1989).

To think about the worst-case payoff, the market maker must first have a payoff function. Recall that the benchmark model in Section IV.3 implicitly assumes that the market maker's goal is to find a pricing strategy that minimizes the mean squared price error (see equation (IV.1)).[12] Thus, with an unknown $\bar{s}$, a natural idea is to assume that the market maker chooses a (linear) pricing strategy $P^r(s, y)$ that minimizes the maximal (across all possible values of $\bar{s}$) mean squared price error:[13]

$$\min_{\tilde{P}(s,y) \text{ is affine}} \max_{\tilde{s} \in \mathbb{R}} \mathbb{E}_{\tilde{s}}[(\tilde{P}(s, y) - v)^2],$$

in which $\mathbb{E}_{\tilde{s}}[\cdot]$ denotes the expectation assuming that the joint distribution of v and s is $\mathcal{N}\left(\begin{bmatrix} \bar{v} \\ \tilde{s} \end{bmatrix}, \begin{bmatrix} \sigma_v^2 & \rho\sigma_v\sigma_s \\ \rho\sigma_v\sigma_s & \sigma_s^2 \end{bmatrix}\right)$, even though the actual mean of s is $\bar{s}$.

An obvious problem with this idea is whether it is reasonable to assume that the market maker's payoff is well described by the mean squared function. Will his behavior be rather different if we assume instead that his payoff function is the

[12] Kyle (1985) offers a second interpretation of the market maker's behavior: At least two risk-neutral profit-maxmizing market makers simultaneously compete on prices; that is, traders' orders go to the market maker with the best price. Such a Bertrand competition drives market makers' profits to zero and leads to condition 1 of Definition IV.1. The robust version of such a Bertrand competition, however, is nontrivial and beyond the scope of this paper.

[13] We allow the value of the objective function to be $\pm\infty$ (extended real numbers) in all minimax problems in the paper.

absolute difference between $\tilde{P}(s,y)$ and v?

Therefore, we consider a more general robustness problem for the market maker. We assume that the market maker minimizes the maximal (across all possible values of $\bar{s}$) price error under some *loss function* c:

$$\min_{\tilde{P}(s,y) \text{ is affine}} \max_{\bar{s}\in\mathbb{R}} \mathbb{E}_{\bar{s}}[c(|\tilde{P}(s,y)-v|)]. \tag{IV.2}$$

Then, at a time when v is revealed after $t=1$, the probabilistically informed trader receives profit $\pi = (v - P^r(s,y))X(\bar{s},s)$. We call a function $c:\mathbb{R}_+\to\mathbb{R}_+$ a loss function if c is twice continuously differentiable, $c(0)=0$, $c'\geq 0$, $c''\geq 0$, and $c''=0 \Rightarrow c'>0$.[14] The last condition is to rule out the case in which c is a zero function. A straightforward and useful implication of these assumptions is that c is unbounded. A loss function describes the market maker's attitude toward mispricing.

While (IV.2) is general in the sense that the loss function is arbitrary, it makes three restrictive assumptions. First, it is assumed that the underlying variable the market maker cares about is the price error. This assumption is taken from Definition IV.2 and is a simple natural starting point. Our results will depend on it—if, for example, the market maker instead cares about the trading volume, results could be rather different.[15] Second, we focus on linear trading and pricing strategies. Again, our results crucially depend on this assumption. Moreover, although this assumption is taken from Kyle (1985), there is some difference. In Kyle, the assumption that the insider's trading strategy is linear implies that the market maker's optimal pricing strategy is linear, but this is not necessarily the case in our model. Last, we assume that the market maker faces a special type of ambiguity. On the one hand, the joint distributions the market maker considers plausible only differ in one parameter, the

[14] We can define c on $\mathbb{R}$, but we must then allow c to not be differentiable at 0 to allow the loss function to be the absolute value function.

[15] One way to understand this assumption is to think of the market maker as someone hired to clear the market and set the price. Suppose that the main goal of the market maker's employer (some firm, or perhaps the government) is to ensure that its clients—traders who might be paying a fixed amount of money (not modeled explicitly in our setup) to the market maker's employer—feel that prices are set "fairly" so that they are willing to continue to be its clients. Then, the employer may want to reward or punish the market maker based on the price error.

mean of s. On the other hand, the market maker believes that this parameter may take any value; that is, this parameter cannot be bounded above or below. This assumption is important in making (IV.2) tractable.

Next, we define the equilibrium. The definition below is a robust (linear) version of Definition IV.2.

Definition IV.3. *The pair of affine functions $X(\bar{s}, s)$ and $P^r(s, y)$ is a linear robust pricing (RP) equilibrium if*

1. *given $X(\bar{s}, s)$, $P^r(s, y)$ solves (IV.2); and*
2. *given $P^r(s, y)$, $X(\bar{s}, s)$ maximizes $\mathbb{E}_{\bar{s}}[\pi|s]$.*

Before we analyze the linear RP equilibrium, let us emphasize that the market maker is non-Bayesian. He does not have a prior over $\bar{s}$. We believe that this is a better description of the situation we want to model. Nonetheless, there will be a connection between our approach and the Bayesian approach. We will return to this at the end of this section and compare the two approaches more formally in Section 7.1.

The solution to the market maker's robustness problem has a simple characterization, which will shed light on a basic question about the robustness approach. Without knowing the joint distribution, the market maker faces ambiguity. What does the market maker do with ambiguity? For example, does the market maker "learn" the distribution as he implements the robust pricing strategy? This question will be particularly relevant in the dynamic model, and our model provides a sharp answer.

IV.4.1 An Equivalent Two-step Learning Procedure

Our first main result will show that assuming that the market maker solves the robustness problem (IV.2) is equivalent to assuming that he follows the following two-step learning procedure. The market maker first estimates the unknown parameter using the available data, which is what people usually do in practice when they

face unknown distributions. He then updates his belief about v according to the estimated distribution and public signal s.

Specifically, after observing s and y, the market maker first computes an estimate of $\bar{s}$, denoted by $\hat{s}(s,y)$. There are many ways to estimate $\bar{s}$ in general, such as the maximum likelihood method, ordinary least squares, generalized method of moments, etc., but a basic rationality requirement will suggest that we assume the market maker uses the "optimal" method to estimate $\bar{s}$. Then, the market maker uses the estimated joint distribution, $\mathcal{N}\left(\begin{bmatrix} \bar{v} \\ \hat{s}(s,y) \end{bmatrix}, \begin{bmatrix} \sigma_v^2 & \rho\sigma_v\sigma_s \\ \rho\sigma_v\sigma_s & \sigma_s^2 \end{bmatrix}\right)$, to determine the price at which orders are traded, $P_{\hat{s}(s,y)}(s,y)$, according to the conditional expectation of v.

The notion of the optimal estimator is standard in econometrics and statistics. Again, similar to Kyle (1985), we restrict our attention to a linear setting.

Definition IV.4. *A real-valued function $\hat{s}(s,y)$ is an unbiased estimator of $\bar{s}$ if $\mathbb{E}[\hat{s}(s,y)] = \bar{s}$, and a linear estimator if it is affine.*[16] *A real-valued function $\hat{s}(s,y)$ is a best linear unbiased estimator (BLUE) of $\bar{s}$ if (i) $\hat{s}(s,y)$ is a linear unbiased estimator of $\bar{s}$ and (ii) the variance of $\hat{s}(s,y)$ is the lowest among all linear unbiased estimators of $\bar{s}$.*

The probabilistically informed trader's strategy determines the information content of y, and hence affects whether an estimator of $\bar{s}$ is unbiased or minimum-variance.

The result below shows that when the market maker adopts a robust linear pricing strategy, it is as if he learns $\bar{s}$ as well as he can following the two-step learning procedure. From here on, all omitted proofs can be found in the Appendix.

Proposition IV.2. *Given any affine $X(\bar{s},s)$, $P^r(s,y) = \mathbb{E}_{\hat{s}(s,y)}[v|s,y]$, in which $P^r(s,y)$ solves (IV.2) and $\hat{s}(s,y)$ is the unique BLUE of $\bar{s}$.*

This characterization is related to the literature on minimax estimation. If we require that the market maker's loss function be the squared function, Proposition

[16] In other words, $\hat{s}(s,y) = k_1 + k_2 s + k_3 y$ for some constants k_1, k_2, k_3.

IV.2 would follow from some classic results in Chapter 5 of Lehmann and Casella (1998) and Chapter 4 of Shao (2003) applied to our setting. Similar to our model, Hirano and Porter (2003a,b) consider general loss functions; they discuss the equivalence between the minimax estimator and the maximum likelihood estimator. We will soon show that the maximum likelihood method is one way to derive the BLUE of $\bar{s}$ in our model, and therefore Proposition IV.2 is closely related to Hirano and Porter's findings. The loss function Hirano and Porter (2003b) consider does not need to be differentiable and hence is more general than our result if applied to our setting. Our proof, however, uses calculus and is to some extent simpler. In addition, in the dynamic model, dynamic programming will turn out to be important for a similar characterization result to hold, which does not appear in the results mentioned above.

Proposition IV.2 shows that regardless of the loss function, the solution to the market maker's robustness problem (IV.2) is the same: First, set the worst-case mean of the price error to zero and then minimize its variance. To see this, take an arbitrary linear trading strategy $X(\bar{s}, s) = \alpha_1 + \alpha_2 s + \alpha_3 \bar{s}$. The market maker's pricing strategy $P^r(s, y)$ should solve the robustness problem:

$$\min_{\tilde{P}(s,y) \text{ is affine}} \max_{\tilde{s} \in \mathbb{R}} \mathbb{E}_{\tilde{s}}[c(|\tilde{P}(s, y) - v|)].$$

Let $\tilde{P}(s, y) = \lambda_1 + \lambda_2 s + \lambda_3 y$. Since $y = X(\bar{s}, s) + u$,

$$v - \tilde{P}(s, y) = v - \lambda_1 - \lambda_2 s - \lambda_3(\alpha_1 + \alpha_2 s + \alpha_3 \tilde{s} + u),$$

which is a normal random variable. Its mean is $\mu = \bar{v} - \lambda_1 - \lambda_3\alpha_1 - (\lambda_2 + \lambda_3\alpha_2 + \lambda_3\alpha_3)\tilde{s}$ and variance is $\sigma^2 = \sigma_v^2 + (\lambda_2 + \lambda_3\alpha_2)^2\sigma_s^2 + \lambda_3^2\sigma_u^2 - 2(\lambda_2 + \lambda_3\alpha_2)\rho\sigma_v\sigma_s$.

Note that the mean μ is affine in $\tilde{s}$, but the variance σ^2 is independent of $\tilde{s}$. Because the market maker will consider the worst $\tilde{s}$, if the coefficient of $\tilde{s}$ in μ is nonzero, his worst-case expected loss will be infinite. Thus $\lambda_2 + \lambda_3\alpha_2 + \lambda_3\alpha_3 = 0$.

In fact, μ itself must be zero. Define $f(\mu, \sigma) := \mathbb{E}[c(|\tilde{P}(s, y) - v|)]$, taking into account $\lambda_2 + \lambda_3\alpha_2 + \lambda_3\alpha_3 = 0$. We first verify that under our assumptions on

c, $\frac{\partial f}{\partial \mu}\Big|_{\mu=0} = 0$ and $\frac{\partial^2 f}{\partial \mu^2} > 0$. Next, suppose the solution of the robustness problem $(\lambda_1^\star, \lambda_2^\star, \lambda_3^\star)$ implies that the mean of $\tilde{P}(s,y)-v$ is $\mu^\star \neq 0$ and the standard deviation of $\tilde{P}(s,y)-v$ is $\sigma^\star$. Since $\sigma^\star$ does not depend on $\lambda_1^\star$, we can replace $\lambda_1^\star$ with $\tilde{\lambda}_1^\star = \bar{v}-\lambda_3^\star\alpha_1$. This change does not affect $\sigma^\star$ and sets the mean of $\tilde{P}(s,y) - v$ to zero. Since $\frac{\partial f}{\partial \mu}\Big|_{\mu=0} = 0$ and $\frac{\partial^2 f}{\partial \mu^2} > 0$, this change must have reduced the expected loss. Therefore, the mean of $\tilde{P}(s,y) - v$ must be zero.

Finally, we verify that $\frac{\partial f}{\partial \sigma}\big|_{\mu=0} > 0$. This means that to solve the robustness problem, we only need to ensure that the mean of $\tilde{P}(s,y) - v$ is zero and minimize its variance, which eventually implies that the BLUE of $\bar{s}$ must be used.

Although we have only considered normal distributions so far, the above proof strategy continues to work if we assume that the joint distribution of v and s is instead elliptical. Normal distributions are elliptical, and some elliptical distributions are thick-tailed. In the Appendix, we explain how to generalize Proposition IV.2 to the case with elliptical distributions.

One may wonder what will happen if we assume instead that the market maker does not know ρ. There are two difficulties. First, when $\bar{s}$ is unknown, the market maker has two sources of information to learn $\bar{s}$: s and y. When ρ is unknown, the market maker will only have one source of information (y) to learn ρ, in which case the strategic interaction is different—it turns out that the market maker must ignore y when setting the price in equilibrium. One natural idea for fixing this is to give the market maker an additional noisy signal of ρ. However, because whenever ρ appears in the strategy it is always multiplied with s, the problem often becomes nonlinear. For example, assuming that the noisy signal of ρ is uniformly distributed between 0 and 1 renders the problem intractable. We may assume that the noisy signal of ρ follows a symmetric two-point discrete distribution to regain tractability, but this distribution seems unrealistic and no longer gives us the equivalence between the robustness problem and the two-step learning procedure.

Next, we define a useful alternative equilibrium based on the two-step learning procedure, which modifies the market maker's objective in Definition IV.1. Under the notations for the two-step learning procedure, at the time when v is revealed

after $t = 1$, the profit the probabilistically informed trader receives is $\pi = (v - P_{\hat{s}(s,y)}(s,y))X(\bar{s},s)$.

Definition IV.5. *The pair of affine functions $X(\bar{s},s)$ and $P_{\hat{s}(s,y)}(s,y)$ is a BLUE equilibrium if*

1. *given $X(\bar{s},s)$, $P_{\hat{s}(s,y)}(s,y) = \mathbb{E}_{\hat{s}(s,y)}[v|s,y]$, in which $\hat{s}(s,y)$ is a BLUE of $\bar{s}$; and*
2. *given $\hat{s}(s,y)$ and $P_{\hat{s}(s,y)}(s,y)$, $X(\bar{s},s)$ maximizes $\mathbb{E}_{\bar{s}}[\pi|s]$.*

In the benchmark model (and most models that follow Kyle (1985)), we can think of the market maker's goal as to offer a reasonable price (Definition IV.1) or, equivalently, as the result of minimizing the mean squared price error (Definition IV.2). The BLUE equilibrium extends Definition IV.1 and the linear RP equilibrium extends Definition IV.2.

Observe that Proposition IV.2 holds for an arbitrary affine $X(\bar{s},s)$, which is not necessarily an equilibrium trading strategy. Then, together with the observation that the probabilistically informed trader's behavior is identical in Definitions IV.3 and IV.5, Proposition IV.2 implies that all market participants' best responses in a linear RP equilibrium are the same as those in a BLUE equilibrium. The following corollary immediately follows.

Corollary IV.1. *Every linear RP equilibrium is a BLUE equilibrium, and vice versa.*

IV.4.2 Characterization of the Equilibrium

We characterize the linear RP equilibrium indirectly through the BLUE equilibrium. The comparison between (the dynamic generalization of) this equilibrium and the benchmark model will be discussed in Section 6.

Theorem IV.1. *There exists a unique BLUE equilibrium in which $\hat{s}(s,y) = s - \frac{\sigma_s}{2\sigma_u}y$, $X(\bar{s},s) = \frac{\sigma_u}{\sigma_s}(s-\bar{s})$, and $P_{\hat{s}(s,y)}(s,y) = \bar{v} + \frac{\rho\sigma_v}{2\sigma_u}y$.*

Due to Proposition IV.2, we know that the unique linear RP equilibrium must consist of $X(\bar{s},s) = \frac{\sigma_u}{\sigma_s}(s-\bar{s})$ and $P^r(s,y) = \bar{v} + \frac{\rho\sigma_v}{2\sigma_u}y$. We characterize the linear

RP equilibrium indirectly, because the BLUE equilibrium offers additional procedural interpretation of the market maker's behavior. In the dynamic model, the estimation of $\bar{s}$ from the indirect characterization will play a more important role.

Without knowing $\bar{s}$, when observing the public signal s, the market maker does not know whether s should be interpreted as a positive or negative signal. If the market maker observes a high total order y, however, he will infer that the public signal s is most likely positive; that is, $\bar{s}$ should be well below s, which means that $\hat{s}(s, y)$ should be decreasing in y.

The equilibrium pricing strategy does not depend on s directly. The intuition is that s alone cannot help the market maker determine what price yields a smaller price error, but the total order y can. It is equal to $\frac{\sigma_u}{\sigma_s}(s - \bar{s}) + u$ in equilibrium, in which the difference term $s - \bar{s}$ is what matters in predicting the value of the asset v. Therefore, the market maker only needs y to set the optimal price in equilibrium.

How exactly does the market maker estimate $\bar{s}$? Below, we discuss two methods often used by economists and explain how they are applied to our setting to generate the unique BLUE of $\bar{s}$: the *maximum likelihood method* and the *optimal generalized method of moments* (optimal GMM).[17] We focus on the latter, since it is useful in the proof of the dynamic model.

IV.4.2.1 The BLUE of $\bar{s}$

A GMM estimator uses moment conditions and a weighting matrix. To apply the GMM in our setting, we first analyze the probabilistically informed trader's behavior. The probabilistically informed trader faces the profit maximization problem

$$\max_X \mathbb{E}_{\bar{s}}[(v - P_{\hat{s}(s,y)}(s, y))X|s]. \tag{IV.3}$$

Suppose she believes that the market maker's linear pricing strategy is

$$P_{\hat{s}(s,y)}(s, y) = \lambda_1 + \lambda_2 s + \lambda_3 y. \tag{IV.4}$$

[17] See Hansen (1982), Greene (2012), and Wooldridge (2016).

Then,

$$\begin{aligned}\mathbb{E}_{\bar{s}}[(v - P_{\hat{s}(s,y)}(s,y))X|s] &= \mathbb{E}_{\bar{s}}[(v - \lambda_1 - \lambda_2 s - \lambda_3(X+u))X|s] \\ &= -\lambda_3 X^2 + (\mathbb{E}_{\bar{s}}[v|s] - \lambda_1 - \lambda_2 s)\, X,\end{aligned}$$

and the solution to (IV.3) given (IV.4) is

$$X(\bar{s}, s) = \frac{1}{2\lambda_3}\left[\bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}) - \lambda_1 - \lambda_2 s\right]. \tag{IV.5}$$

The moment conditions come from $\mathbb{E}(s) = \bar{s}$ and $\mathbb{E}(y|s) = X(\bar{s}, s)$ (because $y = X(\bar{s}, s)+u$). A GMM estimator with an arbitrary positive-definite 2×2 weighting matrix W is the solution $\hat{s}_W(s, y)$ to a minimization problem:

$$\hat{s}_W(s,y) = \arg\min_{\tilde{s}} \left(\begin{bmatrix} s \\ y \end{bmatrix} - \begin{bmatrix} \tilde{s} \\ X(\tilde{s}, s) \end{bmatrix}\right)' W \left(\begin{bmatrix} s \\ y \end{bmatrix} - \begin{bmatrix} \tilde{s} \\ X(\tilde{s}, s) \end{bmatrix}\right); \tag{IV.6}$$

that is, given s, y, and W, the GMM estimator minimizes the deviation from the two moment conditions. Let $\mathbb{W}$ be the set of all positive-definite 2×2 matrices. When a GMM estimator $\hat{s}_W(s, y)$ satisfies

$$\mathrm{Var}(\hat{s}_W(s,y)) = \min_{\tilde{W}\in\mathbb{W}} \mathrm{Var}(\hat{s}_{\tilde{W}}(s,y)),$$

it is called an optimal GMM estimator. Let us use $\hat{s}_{GMM}(s, y)$ to denote an optimal GMM estimator. The following lemma follows from standard econometrics/statistics arguments.

Lemma IV.1. *The unique BLUE $\hat{s}(s, y)$ of $\bar{s}$ satisfies $\hat{s}(s, y) = \hat{s}_{GMM}(s, y)$.*

The lemma does not say that to obtain a BLUE of $\bar{s}$ the market maker must use the optimal GMM. It only says that if an estimator of $\bar{s}$ is a BLUE, it will be equal to $\hat{s}_{GMM}(s, y)$. Indeed, the next lemma confirms that the maximum likelihood method yields the same estimator, which also follows from standard econometrics/statistics arguments. The maximum likelihood estimator $\hat{s}_{ML}(s, y)$ induces a joint distribution

$\mathcal{N}\left(\begin{bmatrix} \bar{v} \\ \hat{s}_{ML}(s,y) \end{bmatrix}, \begin{bmatrix} \sigma_v^2 & \rho\sigma_v\sigma_s \\ \rho\sigma_v\sigma_s & \sigma_s^2 \end{bmatrix}\right)$ that maximizes the likelihood of observing s and y.

Lemma IV.2. $\hat{s}_{ML}(s,y) = \hat{s}_{GMM}(s,y)$.

Not all popular econometric/statistical methods yield a BLUE of $\bar{s}$. Obviously, if the market maker uses a nonoptimal GMM estimator, the estimator may not be a BLUE of $\bar{s}$. In fact, if the market maker uses some estimator that is not a BLUE, there may exist infinitely many equilibria. We discuss this in detail in Section IV.7.2.

Finally, it is well known that the maximum likelihood method is often equivalent to a Bayesian approach with an improper uniform prior. Indeed, in addition to the two-step learning procedure characterization in Proposition IV.2, our robustness approach is equivalent to a Bayesian approach in which the market maker has an improper uniform prior over $\bar{s}$ (see the discussion after Proposition IV.5). Because this additional characterization of the robustness approach does not offer any new insight beyond the two-step learning procedure characterization, and does not separate the learning of $\bar{s}$ and the prediction of v as clearly as in the two-step learning procedure, we will focus on the latter. The estimators of $\bar{s}$ emphasized in the two-step learning procedure will be important in the dynamic model.

IV.5 The Dynamic Model

Does the characterization result in the static model (Proposition IV.2) continue to hold in a dynamic setting? If yes, what does the equilibrium look like, and will the market maker eventually "learn $\bar{s}$" as he implements the robust pricing strategy? The dynamic model addresses these questions.

The basic assumptions of the setup are the same as in Section IV.2. Following the public information disclosure, there are N trading dates, $0 < t_1 < t_2 < \cdots < t_N = 1$. Let $\Delta t_n = t_n - t_{n-1}$, in which $t_0 = 0$. At each t_n, the probabilistically informed trader's order is $X_n(\bar{s}, s, p_1, \ldots, p_{n-1})$, in which $p_1, \ldots, p_{n-1}$ denote the past prices she observes. Liquidity traders' order is $u_n \sim \mathcal{N}(0, \sigma_u^2 \Delta t_n)$ and independent of all

other random variables. Let $y_n = X_n(\bar{s}, s, p_1, \ldots, p_{n-1}) + u_n$ be the total order at t_n. Given the current and past total orders, the market maker determines the price $P_n^r(s, y_1, \ldots, y_n)$ to minimize the maximal sum of current and future price errors under some loss function. We use $\mathbf{X}$ and $\mathbf{P}^r$ to denote the N-tuples $(X_1, \ldots, X_N)$ and $(P_1^r, \ldots, P_N^r)$, respectively. When v is revealed after $t = 1$, let $\pi_n = \sum_{m=n}^{N}[v - P_m^r]X_m$ denote the profit the probabilistically informed trader receives for her orders submitted at $t_n, \ldots, t_N$. Sometimes, to emphasize that π_n depends on $\mathbf{X}$ and $\mathbf{P}^r$, we write $\pi_n(\mathbf{X}, \mathbf{P}^r)$. Below, we define the equilibrium.

Definition IV.6. *The set of affine functions $\{X_m, P_m^r\}_{m=1}^N$ is a dynamic linear RP equilibrium if for each $n \in \{1, \ldots, N\}$,*

1. *given $\mathbf{X}$ and $\{P_m^r\}_{m=1}^{n-1}$, $P_n^r, \ldots, P_N^r$ solve*

$$\min_{\tilde{P}_n, \ldots, \tilde{P}_N \text{ are affine}} \max_{\tilde{s} \in \mathbb{R}} \mathbb{E}_{\tilde{s}} \left[\sum_{l=n}^{N} c(|\tilde{P}_l(s, y_1, \ldots, y_l) - v|) \right]; \tag{IV.7}$$

   and

2. *given $\mathbf{P}^r$ and $\{X_m\}_{m=1}^{n-1}$, $X_n, \ldots, X_N$ solve*

$$\max_{\tilde{X}_n, \ldots, \tilde{X}_N \text{ are affine}} \mathbb{E}_{\bar{s}}[\pi_n((X_1, \ldots, X_{n-1}, \tilde{X}_n, \ldots, \tilde{X}_N), \mathbf{P}^r) | s, p_1, \ldots, p_{n-1}].$$

A dynamic linear RP equilibrium is recursive if $P_n^r = P_{n-1}^r + \lambda_n y_n$ for $n = 1, \ldots, N$ and some constants $\lambda_1, \ldots, \lambda_N$.

Discounting can easily be added to the definition of π_n and the market maker's loss function, but is not essential to our results and is ignored for simplicity. Throughout the paper, it is understood that the pricing strategy at time zero is to let the price be $\bar{v}$.

Under ambiguity, there are two popular approaches to model learning/updating: the full Bayesian (prior-by-prior) approach and the maximum likelihood approach (see Cheng (2020) for a recent discussion). The robustness problem in (IV.7) implies

that we take the full Bayesian approach. Similar to many other models of ambiguity (e.g., Hansen and Sargent (2001)), our market maker's preference over linear pricing strategies induced by (IV.7) does not satisfy the dynamic consistency axiom and the rectangularity condition of Epstein and Schneider (2003). However, the market maker is sophisticated: On each trading date, he correctly anticipates his future pricing strategies.

The following result shows how Proposition IV.2 can be extended to the dynamic setting.

Theorem IV.2. *Suppose the set of affine functions* $\{X_m, P_m^r\}_{m=1}^N$ *satisfies part 1 of Definition IV.6 for each* $n \in \{1, \ldots, N\}$. *Then, for each* $n \in \{1, \ldots, N\}$, $P_n^r = \mathbb{E}_{\hat{s}_n(s,y_1,\ldots,y_n)}[v|s, y_1, \ldots, y_n]$, *in which* $\hat{s}_n(s, y_1, \ldots, y_n)$ *is the unique BLUE of* $\bar{s}$.

The BLUE of $\bar{s}$ at t_n is defined in the same way as in Definition IV.4, except that an estimate of $\bar{s}$ at t_n is now a function of $s, y_1, \ldots, y_n$.

Theorem IV.2 shows that given any trading strategy, the market maker's optimal robust pricing strategy is equivalent to the two-step learning procedure. If the market maker's loss function at t_n is simply $c(|\tilde{P}_l(s, y_1, \ldots, y_l) - v|)$, Theorem IV.2 will be a straightforward extension to its static version. The market maker's objective function, however, is the sum of current and future price errors measured by the loss function c; he is not myopic.

The assumption that Part 1 of Definition IV.6 holds for each $n \in \{1, \ldots, N\}$ ensures that we can extend Proposition IV.2 to the dynamic setting. This assumption implies that given any trading strategy $\mathbf{X}$, the market maker determines the robust pricing strategy $\mathbf{P}^r$ through backward induction—at t_N, P_N^r must be optimal; at t_{N-1}, given that the optimal P_N^r will be used on the next trading date, P_{N-1}^r must be optimal; and so on.

To see why backward induction is important, consider an alternative assumption. Suppose the theorem only requires that given an arbitrary $\{X_m, P_m^r\}_{m=1}^N \backslash \{P_n^r\}$, P_n^r minimizes the maximal sum of price errors under c. For example, let $N = 2$. Suppose that at t_1, the market maker determines the optimal P_1^r taking some future pricing strategy P_2^r as given. Note that in this alternative assumption, P_2^r may not be

optimal at t_2. For instance, let P_2^r be equal to $-P_1^r$. Then, the solution to (IV.7) at t_1 may not involve the BLUE of $\bar{s}$.

Under backward induction, in contrast, we can first verify that the optimal robust pricing strategy at t_N is equivalent to the two-step learning procedure using the BLUE of $\bar{s}$. This is an extension of Proposition IV.2. Next, we show that P_{N-1}^r should also be equivalent to the two-step learning procedure using the BLUE of $\bar{s}$. To show this, the key step is to prove that the market maker will not benefit from manipulating the price at t_{N-1} to affect y_N, and hence his payoff $c(|\tilde{P}_{N-1}(s, y_1, \ldots, y_{N-1}) - v|) + c(|\tilde{P}_N(s, y_1, \ldots, y_N) - v|)$.

Imagine that the market maker manipulates the price at t_{N-1}. This affects the probabilistically informed trader's belief and y_N, but it does not affect the market maker's pricing strategy at t_N—the market maker can "undo" the manipulation and convert a manipulated y_N into an unmanipulated one. One can show that fixing $\mathbf{X}$, the information content of y_N is the same regardless of whether p_{N-1} is manipulated or not. This helps us establish that the optimal robust pricing strategy at t_N is independent of the pricing strategy at t_{N-1}. Then, another simple extension of Proposition IV.2 will imply that P_{N-1}^r is also equivalent to the two-step learning procedure using the BLUE of $\bar{s}$.

Again, to obtain a better understanding of the dynamic linear RP equilibrium, we will characterize it indirectly via the equivalent two-step learning procedure. Specifically, at each t_n, based on the public signal s received at $t = 0$ and the orders $y_1, \ldots, y_n$, the market maker computes the BLUE of $\bar{s}$, denoted by $\hat{s}_n(s, y_1, \ldots, y_n)$. Next, using the estimated joint distribution $\mathcal{N}\left(\begin{bmatrix} \bar{v} \\ \hat{s}_n \end{bmatrix}, \begin{bmatrix} \sigma_v^2 & \rho\sigma_v\sigma_s \\ \rho\sigma_v\sigma_s & \sigma_s^2 \end{bmatrix}\right)$, the market maker determines the price, $P_{n,\hat{s}_n}(s, y_1, \ldots, y_n)$, at which the orders at t_n are traded according to the conditional expectation of v. We use $\mathbf{P}_{\hat{s}}$ to denote the N-tuple $(P_{1,\hat{s}_1}, \ldots, P_{N,\hat{s}_N})$. Under these notations for the two-step learning procedure, when v is revealed after $t = 1$, the profit the probabilistically informed trader receives for her orders at $t_n, \ldots, t_N$ is $\pi_n = \sum_{m=n}^N [v - P_{m,\hat{s}_m}]X_m$. Sometimes, to emphasize the fact that π_n depends on $\mathbf{X}$ and $\mathbf{P}_{\hat{s}}$, we write $\pi_n(\mathbf{X}, \mathbf{P}_{\hat{s}})$.

Definition IV.7. *The set of affine functions $\{X_m, P_{m,\hat{s}_m}\}_{m=1}^N$ is a dynamic BLUE*

equilibrium if for each $n = 1, \ldots, N$,

1. *given* $\mathbf{X}$ *and* $\{P_{m,\hat{s}_m}\}_{m=1}^{n-1}$, $P_{n,\hat{s}_n} = \mathbb{E}_{\hat{s}_n}[v|s, y_1, \ldots, y_n]$, *in which* $\hat{s}_n$ *is a BLUE of* $\bar{s}$; *and*

2. *given* $\mathbf{P}_{\hat{s}}$ *and* $\{X_m\}_{m=1}^{n-1}$, $X_n, \ldots, X_N$ *solve*

$$\max_{\tilde{X}_n, \ldots, \tilde{X}_N \text{ are affine}} \mathbb{E}_{\bar{s}}[\pi_n((X_1, \ldots, X_{n-1}, \tilde{X}_n, \ldots, \tilde{X}_N), \mathbf{P}_{\hat{s}})|s, p_1, \ldots, p_{n-1}].$$

A dynamic BLUE equilibrium is recursive if $P_{n,\hat{s}_n} = P_{n-1,\hat{s}_{n-1}} + \lambda_n y_n$ *for* $n = 1, \ldots, N$ *and some constants* $\lambda_1, \ldots, \lambda_N$.

It is implicit that $\hat{s}_0 = s$, because before the market maker observes any order, the BLUE of $\bar{s}$ is equal to the public signal s. Due to Theorem IV.2 and the fact that the probabilistically informed trader's behavior is identical in Definitions IV.6 and IV.7, we know that all market participants' *backward-induction* best responses in a dynamic linear RP equilibrium are the same as those in a dynamic BLUE equilibrium. Hence, we have the following corollary.

Corollary IV.2. *Every dynamic linear RP equilibrium is a dynamic BLUE equilibrium, and vice versa.*

Again, we characterize the dynamic linear RP equilibrium indirectly through the dynamic BLUE equilibrium, and leave the comparison between this equilibrium and the benchmark model to Section 6.

Theorem IV.3. *There exists a unique dynamic BLUE equilibrium and it is recursive. In the dynamic BLUE equilibrium,*

$$\begin{aligned}
\hat{s}_n &= \hat{s}_{n-1} - \frac{\sigma_s}{\rho\sigma_v}\lambda_n y_n, \\
P_{n,\hat{s}_n} &= P_{n-1,\hat{s}_{n-1}} + \lambda_n y_n, \\
X_n &= \beta_n \Delta t_n (\hat{s}_{n-1} - \bar{s}), \\
\mathbb{E}_{\bar{s}}[\pi_n|s, p_1, \ldots, p_{n-1}] &= \alpha_{n-1}(\hat{s}_{n-1} - \bar{s})^2 + \delta_{n-1}, \\
\omega_n &= \omega_{n-1} + \frac{\beta_n^2 \Delta t_n}{\sigma_u^2}
\end{aligned}$$

in which $\{\alpha_n, \beta_n, \lambda_n, \delta_n, \omega_n\}_{n=1}^N$ *are the unique solution to the following difference equation system:*

$$\begin{aligned}
\lambda_n \beta_n \Delta t_n &= \frac{\rho \sigma_v}{\sigma_s}\left(1 - \frac{\omega_{n-1}}{\omega_n}\right), \\
\frac{\omega_{n-1}}{\omega_n} &= \frac{1}{2\left(1 - \frac{\alpha_n \lambda_n}{(\rho\sigma_v/\sigma_s)^2}\right)}, \\
\alpha_{n-1} &= \left(\frac{\omega_{n-1}}{\omega_n}\right)^2 \alpha_n + \frac{\rho \sigma_v}{\sigma_s}\frac{\omega_{n-1}}{\omega_n}\beta_n \Delta t_n, \\
\delta_{n-1} &= \delta_n + \frac{\alpha_n \beta_n^2 \Delta t_n}{\omega_n^2 \sigma_u^2}
\end{aligned}$$

subject to $\omega_0 = 1/\sigma_s^2$, $\alpha_N = \delta_N = 0$, *and* $\lambda_n\left(1 - \frac{\alpha_n \lambda_n}{(\rho\sigma_v/\sigma_s)^2}\right) > 0$. *Moreover,* $\{\alpha_n, \beta_n, \lambda_n, \delta_n, \omega_n\}_{n=1}^N$ *are nonnegative.*

In the dynamic BLUE equilibrium, if we plug $\hat{s}_n$ into the trading and pricing strategies, we will obtain the unique dynamic linear RP equilibrium. Alternatively, $\hat{s}_n$ may be thought of as an auxiliary variable in the dynamic linear RP equilibrium. This auxiliary variable not only helps us better understand the market maker's behavior, but also the probabilistically informed trader's. By introducing $\hat{s}_n$, we can see that the structure of the dynamic linear RP equilibrium is similar to Kyle's (1985) equilibrium: The equilibrium trading strategy is proportional to the probabilistically informed trader's information advantage, and the current and future profit term π_n is a quadratic function of the information advantage. This is essentially due to the linearity, normal distribution, and risk neutrality assumptions.

The proof strategy of Theorem IV.3 differs from Kyle's (1985). For example, in Kyle, the martingale property of the prices is crucial and directly implies that prices are recursive. This greatly simplifies the proof of Kyle's result. The same proof idea does not go through in our model with an unknown $\bar{s}$. In our model, to show that prices are recursive, we first prove that in a dynamic BLUE equilibrium, at each t_n, the optimal GMM estimator of $\bar{s}$ is equal to the unique BLUE of $\bar{s}$. Next, we show that the optimal GMM estimator can be derived recursively: Although at each t_n

an estimator of $\bar{s}$ can depend on $s, y_1, \ldots, y_{n-1}, y_n$, the market maker only needs to use the new order y_n and the optimal GMM estimator of $\bar{s}$ at t_{n-1}, $\hat{s}_{n-1}$, to form the optimal GMM estimator of $\bar{s}$ at t_n. This is a key step, but is still insufficient to show that prices are recursive. We guess the structure of the equilibrium directly, and verify that prices are recursive as we verify our guess.

IV.6 Underreaction and Market Efficiency

Corollary IV.2 and the dynamic BLUE equilibrium enable us to unveil two interesting properties of the dynamic linear RP equilibrium. First, on average, we observe *underreaction* to public information. Suppose there is a positive public signal (from the probabilistically informed trader's point of view, or equivalently, under the true joint distribution of v and s). At each t_n, the equilibrium price on average reacts less sensitively to the public signal s compared with the benchmark model. Moreover, the equilibrium price on average gradually increases toward $\mathbb{E}_{\bar{s}}[v|s]$, which is the conditional expectation of the value of the asset under the true joint distribution.

However, these statements seem to require that at least to some outside researchers, the true joint distribution must be known, even though the model assumes that the only person who knows $\bar{s}$ is the probabilistically informed trader. The second property of the dynamic BLUE equilibrium addresses this concern, as well as the question about what the market maker will do with the ambiguity about $\bar{s}$ as he solves the robustness problem. We show that as Δt_n's go to zero, the price at $t = 1$, $P_{N,\hat{s}_N}$, converges to $\mathbb{E}_{\bar{s}}[v|s]$; that is, as the market maker implements the robust pricing strategy, it is as if he fully learns $\bar{s}$ in the end if the trading frequency is arbitrarily high. Therefore, if the trading frequency is sufficiently high, the price of the asset will reveal $\bar{s}$ in the end, even though this is not the goal of the market maker in the robustness approach.

Definition IV.8. *We say that a dynamic BLUE equilibrium exhibits underreaction if for each* $n = 1, \ldots, N$,

$$\mathbb{E}_{\bar{s}}[P_{n,\hat{s}_n}(s, y)|s] = \bar{v} + \theta_n \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s})$$

for some $\theta_n \in [0,1)$ *and* $\theta_1 < \theta_2 < \cdots < \theta_N$. *Each* θ_n *is called the underreaction parameter at* t_n.

If $\bar{s}$ is known, the price should be equal to $\bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s})$; that is, the underreaction parameters should be replaced with 1. The result below shows that the dynamic BLUE equilibrium exhibits underreaction.

Proposition IV.3. *The dynamic BLUE equilibrium exhibits underreaction. In particular,* $\theta_n = 1 - \frac{\omega_0}{\omega_n}$.

According to Theorem IV.3, the weights $\omega_0, \omega_1, \ldots, \omega_N$ are positive and strictly increasing. Therefore, $0 < \frac{\omega_0}{\omega_n} < 1$ and $1 - \frac{\omega_0}{\omega_n}$ is strictly increasing. Underreaction is due to the gradual information revelation by the probabilistically informed trader. For a specific realization of the random variables, $s, u_1, \ldots, u_N$, we may not observe that the prices slowly increase (decrease) when the public signal is positive (negative). It could be the case that liquidity traders submit large positive orders early on, and hence the market maker is overly pessimistic about $\bar{s}$. At some later time, the price may fall back to the correct level $\mathbb{E}_{\bar{s}}[v|s]$. In this case, the prices do not monotonically increase.

Knowing the existence of underreaction, why does the market maker not eliminate underreaction if, as in the dynamic BLUE equilibrium, he wants to set a fair price for the asset? [18] It is easier to see the answer if we return to the robustness approach. In the robustness approach, if the market maker uses higher λ_n's (reacts more aggressively to orders compared with the equilibrium), it is not difficult to see that the underreaction parameters will become closer to 1. However, higher λ_n's will render the prices more volatile. According to Theorem IV.2, they will generate higher worst-case price errors measured by *any* loss function c. Therefore, it is not optimal for the market maker to eliminate underreaction.

One may wonder how Proposition IV.3 depends on our assumption about the market maker's objective in the dynamic linear RP equilibrium. For example, is ambiguity or the use of the maxmin expected utility model crucial in deriving the

[18] To be more precise, the market maker knows that there will be underreaction under the true joint distribution of v and s, which is unknown to him.

underreaction result? The answers to these questions are positive and will be provided in Section IV.7.1.

Our theory is not the only one that explains underreaction. A testable implication of our theory that may distinguish it from alternative theories of underreaction, such as Daniel et al. (1998) and Hirshleifer et al. (2009), is as follows. A distinctive feature of our theory of underreaction is that the price drift is rational. One can easily verify that an uninformed arbitrager (a profit-maximizing trader who does not know $\bar{s}$) cannot profit from the price drift in our theory. Therefore, an implication of our theory is that the price drift would not vary with limits to arbitrage such as trading constraints. This prediction is different from alternative behavioral theories that rely on limits to arbitrage to sustain the price drift.

One way to implement this idea is to use the regulation SHO pilot program adopted by the SEC in 2004. Regulation SHO relaxed the trading constraints for a random set of pilot stocks from the Russell 3000 index (see Chu, Hirshleifer, and Ma (2020)). Our theory predicts that the price drift will not differ significantly across pilot firms and non-pilot firms, while alternative behavioral theories predict significant differences.

Next, we examine how much information from the public signal will be absorbed into the price in the end. This depends on how much the market maker eventually learns about $\bar{s}$ in the dynamic BLUE equilibrium. It could be that although θ_n increases as n goes to N, θ_N is still quite far from 1. For example, when $N = 1$, the dynamic model becomes static and $\theta_N = 1/2$, which is quite different from 1. The result below shows that if the traders can trade frequently, the price at $t = 1$, $P_{N,\hat{s}_N}$, will converge to $\mathbb{E}_{\bar{s}}[v|s]$.

Proposition IV.4. *In the dynamic BLUE equilibrium, if Δt_n's go to zero and N goes to infinity, $P_{N,\hat{s}_N}$ converges to $\mathbb{E}_{\bar{s}}[v|s]$.*

Therefore, if the trading frequency is sufficiently high, $\bar{s}$ will be revealed in the price, and underreaction can be verified by anyone ex post. As the market maker solves the dynamic robustness problem, it is as if the ambiguity about $\bar{s}$ is completely eliminated.

This proposition is closely related to Theorems 3 and 4 of Kyle (1985). Kyle shows that as Δt_n's go to zero, his discrete-time model converges to a continuous-time model in which the market maker learns at a constant rate and learns everything eventually. Our proof strategy closely follows that of Kyle's Theorem 4.

IV.7 Assumptions about the Market Maker's Behavior

In this section, we return to the static setting and analyze how our main findings depend on the assumptions about the market maker's behavior. First, we study how our static model differs from a model in which the market maker is Bayesian. The main finding is that underreaction does not always occur when the market maker is Bayesian. Second, we examine what happens when the market maker uses estimators of $\bar{s}$ other than the BLUE in the first step of the two-step learning procedure. We show that there will be multiple equilibria in this case, which makes it possible that the equilibrium trading order and price depend on "animal spirits," as in Epstein and Wang (1994).

IV.7.1 A Bayesian Market Maker

Consider the setup of Section 4. Now, assume that the market maker is Bayesian instead. In particular, the market maker still does not know $\bar{s}$, but he believes that $\bar{s}$ is a random variable that follows the normal distribution $\mathcal{N}(\mu_{\bar{s}}, \sigma_{\bar{s}}^2)$ and is independent of u and v. For each realization of $\bar{s}$, the joint distribution of v and s is the same as before. Thus,

$$\begin{bmatrix} v \\ s \\ \bar{s} \\ u \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \bar{v} \\ \mu_{\bar{s}} \\ \mu_{\bar{s}} \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_v^2 & \rho\sigma_v\sigma_s & 0 & 0 \\ \rho\sigma_v\sigma_s & \sigma_{\bar{s}}^2 + \sigma_s^2 & \sigma_{\bar{s}}^2 & 0 \\ 0 & \sigma_{\bar{s}}^2 & \sigma_{\bar{s}}^2 & 0 \\ 0 & 0 & 0 & \sigma_u^2 \end{bmatrix} \right).$$

Let $P^b(s, y)$ denote the (Bayesian) market maker's pricing strategy. Other notations remain unchanged. Then, we can follow Kyle (1985) to define the following

equilibrium.

Definition IV.9. *The pair of affine functions $X(\bar{s}, s)$ and $P^b(s, y)$ is a linear Bayesian equilibrium if*

1. *given $X(\bar{s}, s)$, $P^b(s, y) = \mathbb{E}[v|s, y]$; and*

2. *given $P^b(s, y)$, $X(\bar{s}, s)$ maximizes $\mathbb{E}[\pi|\bar{s}, s]$.*

Recall that in the benchmark model, we have $P(s, y) = \bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s})$ in equilibrium. From Corollary IV.1 and Theorem IV.1, the static model in Section 4 has $P^r(s, y) = \bar{v} + \frac{\rho\sigma_v}{2\sigma_u}y$ and $X(\bar{s}, s) = \frac{\sigma_u}{\sigma_s}(s - \bar{s})$ in equilibrium. Hence, in equilibrium, viewed under full information (i.e., the probabilistically informed trader's information or, equivalently, the information that will be revealed to the market at the end of all trades in the dynamic model with an infinitely high trading frequency),

$$\mathbb{E}_{\bar{s}}[P^r(s, y)|s] = \bar{v} + \frac{1}{2}\frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}). \tag{IV.8}$$

According to Definition IV.8, 1/2 is the underreaction parameter (on the only trading date). Clearly, for any $\bar{s}$, the expected price always exhibits underreaction regardless of the realization of s. The following result shows that this is not the case when the market maker is Bayesian.

Proposition IV.5. *There exists a unique linear Bayesian equilibrium in which*

$$\begin{aligned}\mathbb{E}[P^b(s, y)|\bar{s}, s] &= \bar{v} + \left(\frac{1}{2} + \frac{\sigma_s^2}{2\left(\sigma_{\bar{s}}^2 + \sigma_s^2\right)}\right)\frac{\rho\sigma_v}{\sigma_s}\left(s - \frac{\bar{s}/\sigma_s^2 + \mu_{\bar{s}}/\left(\sigma_{\bar{s}}^2 + \sigma_s^2\right)}{1/\sigma_s^2 + 1/\left(\sigma_{\bar{s}}^2 + \sigma_s^2\right)}\right) \\ &= \bar{v} + \frac{1}{2}\frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}) + \frac{1}{2}\frac{\rho\sigma_v\sigma_s}{\sigma_{\bar{s}}^2 + \sigma_s^2}(s - \mu_{\bar{s}}).\end{aligned} \tag{IV.9}$$

Let $\check{s}$ denote the weighted average of $\bar{s}$ and $\mu_{\bar{s}}$, $\frac{\bar{s}/\sigma_s^2 + \mu_{\bar{s}}/\left(\sigma_{\bar{s}}^2 + \sigma_s^2\right)}{1/\sigma_s^2 + 1/\left(\sigma_{\bar{s}}^2 + \sigma_s^2\right)}$, in the first equality of (IV.9). To prove Proposition IV.5, we first characterize the unique Bayesian equilibrium in a manner similar to Kyle (1985). It can be seen from the proof that as $\sigma_{\bar{s}}^2$ converges to infinity, the linear Bayesian equilibrium converges to the linear RP equilibrium. In particular, $\mathbb{E}[P^b(s, y)|\bar{s}, s]$ converges to (IV.8).

Focus on the first equality of (IV.9). Compared with (IV.8), the scalar multiplied in front of $\frac{\rho\sigma_v}{\sigma_s}$ is strictly between 1/2 and 1. More importantly, rather than subtracting $\bar{s}$ from s as in (IV.8), $\check{s}$, which depends on $\mu_{\bar{s}}$, is subtracted from s. Therefore, as can be seen from the second equality in Proposition IV.5, we do not always have underreaction under full information (the probabilistically informed trader's information) in the linear Bayesian equilibrium.[19] This suggests that ambiguity is crucial in deriving Proposition IV.3.

This result also suggests that if the maxmin expected utility model in (IV.2) is replaced with another popular model of ambiguity, the smooth ambiguity model (see Klibanoff, Marinacci, and Mukerji (2005)), it is likely that underreaction will not always arise. This is because, similar to the case of the Bayesian market maker, the market maker under the smooth ambiguity model also has a nontrivial expectation of $\bar{s}$.

IV.7.2 "Suboptimal" Estimators of $\bar{s}$

From Section 4, we know that in the two-step learning procedure, the market maker first uses the BLUE to estimate $\bar{s}$. What happens if the market maker uses other estimators? For example, if he uses a biased estimator, how would the equilibrium be affected? Below, we first relax the definition of the BLUE equilibrium so that we can allow the market maker to use other estimators.

Definition IV.10. *The pair of affine functions $X(\bar{s}, s)$ and $P_{\hat{s}(s,y)}(s, y)$ is a linear-estimator (LE) equilibrium if*

1. *given $X(\bar{s}, s)$, $P_{\hat{s}(s,y)}(s, y) = \mathbb{E}_{\hat{s}(s,y)}[v|s, y]$, in which $\hat{s}(s, y)$ is affine; and*

2. *given $\hat{s}(s, y)$ and $P_{\hat{s}(s,y)}(s, y)$, $X(\bar{s}, s)$ maximizes $\mathbb{E}_{\bar{s}}[\pi|s]$.*

If, in addition, $\hat{s}(s, y)$ is unbiased, the pair of functions is a linear-unbiased-estimator (LUE) equilibrium.

[19] Although we do not prove the following claim for the case with a Bayesian market maker, it can be verified that the probabilistically informed trader's information is again the information that will be learned by the market at the end of all trades in the dynamic model with an infinitely high trading frequency.

When we require that the affine function $\hat{s}(s, y)$ be the BLUE of $\bar{s}$, $\hat{s}(s, y)$ needs to satisfy the two assumptions in Definition IV.4. Note that assumption (ii) efficiency implies assumption (i) unbiasedness. Therefore, if we drop both assumptions, we will be interested in LE equilibria. If we impose assumption (ii) only, we will be interested in LUE equilibria.

Proposition IV.6. *The following statements are true:*

1. *For any $\lambda_1, \lambda_2 \in \mathbb{R}$, and $\lambda_3 > 0$, the pair of functions $X(\bar{s}, s) = \frac{1}{2\lambda_3}\left[\bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}) - \lambda_1 - \lambda_2 s\right]$ and $P_{\hat{s}(s,y)} = \lambda_1 + \lambda_2 s + \lambda_3 y$ is an LE equilibrium.*

2. *For any $\lambda_3 > 0$, the pair of functions $X(\bar{s}, s) = \frac{\rho\sigma_v}{2\lambda_3\sigma_s}(s - \bar{s})$ and $P_{\hat{s}(s,y)} = \bar{v} + \lambda_3 y$ is an LUE equilibrium.*

3. *There exist LUE (and hence LE) equilibria such that the probabilistically informed trader's expected profit conditional on s is always lower than that in the BLUE equilibrium regardless of s.*

4. *There exist LUE (and hence LE) equilibria such that the probabilistically informed trader's expected profit conditional on s is always higher than that in the BLUE equilibrium regardless of s.*

Intuitively, the number of properties imposed on the estimator of $\bar{s}$ in the BLUE equilibrium ensures that λ_1, λ_2, and λ_3 can be exactly identified. Once we impose fewer properties, this is no longer the case. The first statement says that if we impose no restriction on the estimator of $\bar{s}$, an arbitrary pricing strategy and the trading strategy that responds to it optimally form an LE equilibrium. The second shows that under unbiasedness, we still have infinitely many LUE equilibria. These LUE equilibria are not trivial equilibria in which the market maker ignores the order and therefore the probabilistically informed trader has infinitely many best responses.

One may wonder if the probabilistically informed trader's expected profit will be higher under LE or LUE equilibria than under the BLUE equilibrium, because the market maker uses "suboptimal" estimators. The third statement says this is not

the case, and the fourth shows that neither is the opposite case true. As for the market maker's payoff, under most LE equilibria, it can be verified that the market maker's maximal expected loss is infinity.

IV.8 Related Literature

A large body of research in macroeconomics and finance has examined situations in which agents face unknown distributions of random variables. On the one hand, similar to our dynamic linear RP equilibrium approach, many studies adopt the maxmin expected utility model to model agents' behavior. For example, Epstein and Wang (1994) show that ambiguity may lead to multiple equilibria in an otherwise standard general equilibrium model. In a continuous-time asset pricing model, Chen and Epstein (2002) characterize an ambiguity premium in addition to the standard risk premium. Epstein and Schneider (2008) find that ambiguity-averse investors react more strongly to bad news than to good news. To study model misspecification, Hansen and Sargent (2008) connect models of ambiguity with robust control techniques and study macroeconomic applications. Several others have followed this approach to study optimal policy design under ambiguity (see Karantounias (2013) and Benigno and Paciello (2014), among others). Easley and O'Hara (2009, 2010) analyze the relation between ambiguity and market participation. Condie and Ganguli (2011) show that under ambiguity, equilibrium market prices may only partially reveal traders' private information even without noise introduced by, for example, liquidity traders. Galanis, Ioannou, and Kotronis (2019) find that information aggregation fails when some traders face ambiguity.

On the other hand, following Hansen (2007), some other studies examine the real-time consequence of distribution estimation, which is related to our dynamic BLUE equilibrium approach. Orlik and Veldkamp (2015) show numerically that if the unknown distribution has non-normal tails, real-time distribution estimation may lead to large uncertainty fluctuations. In Kozlowski, Veldkamp, and Venkateswaran (2019), agents estimate the unknown distribution via the kernel density estimator. It is found numerically that extreme transitory events may induce persistent changes

in beliefs and macroeconomic outcomes.

The maxmin principle is also widely used in the literature on robust contract and mechanism design. Among others, Bose and Renou (2014) show that ambiguous mediated communication in mechanisms helps implement social choice functions that are not incentive compatible. In a moral hazard problem in which the agent may act in ways unknown to the principal, Carroll (2015) shows that the optimal contract is linear. Miao and Rivera (2016) study the optimal contract when the principal faces ambiguity about project cash flows. Di Tillio, Kos, and Messner (2017) show how a seller can use ambiguous mechanisms to increase the profit. Carroll (2017) examines a robust screening problem in which the agent's type is multidimensional, and the principal knows the marginal distribution of each component of the agent's type but not the joint distribution. The optimal solution is that the principal screens along each component separately.

There is a growing literature that studies strategic interactions between players using non-Bayesian econometric/statistical methods. Eliaz and Spiegler (2018) examine incentive compatibility when the principal learns from a penalized regression such as LASSO (see Hastie, Tibshirani, and Wainwright (2015)). In Jéhiel (2018), investors do not know the joint distribution of signals and projects' returns, and implement projects according to a simple heuristic. Duffie and Dworczak (2018) study the estimation of a reference rate such as LIBOR by an administrator who uses the best linear estimator. Liang (2018) analyzes games with players who learn from public information via different methods, and characterizes rationalizable actions and Nash equilibria under a finite dataset. Levy and Razin (2018) consider a dynamic model in which a decision maker adopts the maximum likelihood method to learn from forecasts.

Our paper belongs to the literature on market microstructure. Kyle (1985) shows that in a setting with an insider, liquidity traders, and a market maker, the insider's private information will be slowly incorporated into the price. Following Kyle, many papers have explored other possible information and market structures. For example, Holden and Subrahmanyam (1992) show that the competition among insiders with identical private information may lead to immediate full revelation of private

information. More complicated interactions among insiders with different private information is considered by Foster and Viswanathan (1996) and Back, Cao, and Willard (2000). Huddart, Hughes, and Levine (2001) examine the situation in which the insider needs to disclose their trades. In Caldentey and Stacchetti (2010), the value of the asset is revealed at a random time. Lambert, Ostrovsky, and Panov (2018) characterize the equilibrium in a static setting that allows for arbitrary correlations among the value of the asset, insiders' and market makers' signals, and liquidity traders' orders. In Yang and Zhu (2019), there are fundamental investors and back-runners. The former can observe components of the value of the asset, and the latter have noisy signals about fundamental investors' collective trade.

The closest to us is Hu (2018), who studies a two-period model whose market participants are the insider and the market maker from Kyle (1985), together with a number of arbitrageurs. The distribution of the value of the asset is normal with some probability and Laplacian with some probability. The Laplace distribution has fat tails but the same variance as the normal one. The market maker knows the normal distribution but does not know the possibility of the Laplace distribution. Arbitrageurs receive a common signal about the type of the distribution, but do not know the variance of the Laplace distribution. Hu finds that the arbitrageur's robust optimal strategy is equivalent to learning based on LASSO.

Our paper is also related to the literature on minimax statistical estimation. Under the squared loss function, a partial list includes Wald (1950); Li (1982); Heckman (1988); Lehmann and Casella (1998); and Shao (2003). Under more general loss functions, see Donoho (1994); Hirano and Porter (2003a,b); and Zinodiny, Rezaei, and Nadarajah (2017), as well as our discussion after Proposition IV.2.

It is a well-established empirical finding that prices drift after public events (see footnote 4). One important example is the post-earnings-announcement drift, in which prices continue going up (down) after positive (negative) earnings surprises (see Bernard and Thomas (1989, 1990) and Chan et al. (1996)). Researchers have proposed several explanations for this phenomenon. Daniel et al. (1998) argue that investors are overconfident and hence overweight private information, which causes an underreaction to public information. Hong and Stein (1999) show that two types

of behavioral agents, news-watchers and momentum traders, can together generate not only underreaction in the short run but overreaction in the long run. Hirshleifer et al. (2009) argue that investors do not use full public information due to limited attention. In a recent review, Blankespoor, deHaan, and Marinovic (2020) call for theoretical analyses of how public disclosures may translate into private information due to disclosure processing costs. This is similar to a main idea behind our theory of underreaction: Since different market participants have different abilities to interpret public information, public information generates private information for some participants.

IV.9 Concluding Remarks

We study how people react to public information when the prior distribution is unknown in a model of strategic trading. Market participants consist of liquidity traders, a probabilistically informed trader who knows the joint distribution of public information and an asset's value, and a market maker who does not know the mean of the public signal. We assume that the market maker adopts a robust pricing strategy that induces the best worst-case performance. The performance is defined as the sum of current and future price errors (how much the price deviates from the value of the asset) measured by a general loss function.

We show that the market maker's backward-induction optimal robust pricing strategy is equivalent to the following two-step learning procedure. On each trading date, the market maker first uses the best linear unbiased estimator (BLUE) to estimate the unknown distribution based on public information, market orders, and other participants' strategies. With the estimated distribution, the market maker updates his belief about the value of the asset and, as in Kyle (1985), determines a fair price at which traders' orders are executed.

We characterize the unique linear equilibrium and show that under the true distribution, expected equilibrium prices exhibit underreaction; that is, the price of the asset on average moves in the same direction as the initial impact from the public event for some period of time. Moreover, if the trading frequency is arbitrarily high,

as the market maker solves the dynamic robustness problem, in the end the public information will be fully incorporated into the price and the true distribution will be fully revealed.

APPENDICES

APPENDIX A

Proofs and Discussions for Chapter II

Proof of Theorem II.1. Based on different signs of μ_g and τ_g, the expression of the upper and lower bounds of the closed intervals might be different. However, the critical logic to get the identified set is the same. Here I take the case where $0 \leq \tau_g \leq \mu_g$, for example, and all the other cases can be analyzed based on similar logic.

I know that

$$\begin{aligned}
\mu_g &= \mathbb{E}[g(Y_{i1}(0,1)) - g(Y_{i1}(0,0))|D_i = 1] \\
&= \mathbb{E}[g(Y_{i1}(0,1)) - g(Y_{i0}(0,0)) - (g(Y_{i1}(0,0) - g(Y_{i0}(0,0)))|D_i = 1] \\
&= \mathbb{E}[g(Y_{i1}(1))|D_i = 1] - \mathbb{E}[g(Y_{i0}(0))|D_i = 1] \\
&- (\mathbb{E}[g(Y_{i1}(0))|D_i = 0] - \mathbb{E}[g(Y_{i0}(0))|D_i = 0])
\end{aligned}$$

The first, third and the fourth terms are observable while for the second one

$$\mathbb{E}[g(Y_{i0})|D_i = 1] = \mathbb{E}[g(Y_{i0}(0,1))|D_i = 1] + \mathbb{E}[A_i(g(Y_{i0}(1,1)) - g(Y_{i0}(0,1)))|D_i = 1]$$

$$= \mathbb{E}[g(Y_{i0}(0,1))|D_i = 1] + \mathbb{P}[A_i = 1|D_i = 1]\tau_g \leq \mathbb{E}[g(Y_{i0}(0))|D = 1] + \pi\tau_g$$

For simplicity I call $m_g = \mathbb{E}[g(Y_{i1}) - g(Y_{i0})|D_i = 1] - \mathbb{E}[g(Y_{i1}) - g(Y_{i0})|D_i = 0]$. I

further apply the inequality that $0 \leq \tau_g \leq \mu_g$ and get

$$\tau_g \leq \mu_g \leq m_g + \pi\tau_g$$

. Then I can conclude that

$$0 \leq \tau_g \leq \frac{m_g}{1-\pi} \qquad \tau_g \leq \mu_g \leq \frac{m_g}{1-\pi}$$

From the equation above I notice that $\mu_g = \mathbb{P}[A_i = 1|D_i = 1]\tau_g + m_g \geq m_g$. I can conclude that $\mu \in \left[m_g, \frac{m_g}{1-\pi}\right]$. □

Proof of Theorem II.3. Like in the previous situation, I will take the situation $0 \leq \tau_g \leq \mu_g$ as an example.

$$\begin{aligned}
\mu_g &= \mathbb{E}[g(Y_{i1}(1)) - g(Y_{i1}(0))|D_i = 1] \\
&= \mathbb{E}[g(Y_{i1}(1)) - g(Y_{i0}(0,0)) - (g(Y_{i1}(0) - g(Y_{i0}(0,0)))|D_i = 1] \\
&= \mathbb{E}[g(Y_{i1}(1))|D_i = 1] - \mathbb{E}[g(Y_{i0}(0,0))|D_i = 1] \\
&- (\mathbb{E}[g(Y_{i1}(0))|D_i = 0] - \mathbb{E}[g(Y_{i0}(0,0))|D_i = 0])
\end{aligned}$$

The first and the third term are observable as $\mathbb{E}[g(Y_{i1})|D_i = 1]$ and $\mathbb{E}[g(Y_{i1})|D_i = 0]$. For the second one, those who anticipate their treatment status correctly will make a difference

$$\mathbb{E}[g(Y_{i0})|D_i = 1] = \mathbb{E}[g(Y_{i0}(0,0))|D_i = 1] + \mathbb{P}[A_i = 1|D_i = 1]\tau_g$$

while for the fourth term, those who wrongly anticipate their future will react to it

$$\mathbb{E}[g(Y_{i0})|D_i = 0] = \mathbb{E}[g(Y_{i0}(0,0))|D_i = 0] + \mathbb{P}[A_i = -1|D_i = 0]\tau_g$$

Then I will have

$$\mu_g = \mathbb{E}[g(Y_{i1}(1))|D_i = 1] - \mathbb{E}[g(Y_{i0})|D_i = 1] + \mathbb{P}[A_i = 1|D_i = 1]\tau_g$$

$$-\mathbb{E}[g(Y_{i1}(0))|D_i=0]+\mathbb{E}[g(Y_{i0})|D_i=0]-\mathbb{P}[A_i=-1|D_i=0]\tau_g$$

$$\mu_g = m_g + (\mathbb{P}[A_i\neq 0|D_i=1](1-\varepsilon)-\mathbb{P}[A_i\neq 0|D_i=0]\varepsilon)\tau_g$$

Use the bound from Assumption II.6.4 I can get that the coefficient before τ_g belongs to the interval $[-\pi\varepsilon, \pi(1-\varepsilon)]$ and follow the same approach from the proof of theorem II.1, I will have that $\mu_g \in \left[\frac{m_g}{1+\pi\varepsilon}, \frac{m_g}{1-(1-\varepsilon)\pi}\right]$ □

Including covariates. By taking all the assumptions conditionally, this approach can be generalized to include covariates. We will use notation $b(x)$ to represent the value of parameter b when conditional on "$X=x$". Following similar argument in Theorem II.1, I focus on the case where $0\leq\tau_g(x)\leq\mu_g(x)$. From proof of theorem II.1 I know that

$$\mu_g(x) = \mathbb{E}[g(Y_{i1}(1))|D_i=1,X]-\mathbb{E}[g(Y_{i0}(0))|D_i=1,X]$$

$$-(\mathbb{E}[g(Y_{i1}(0))|D_i=0,X]-\mathbb{E}[g(Y_{i0}(0))|D_i=0,X])$$

The first, third and the fourth term are observable while for the second one

$$\begin{aligned}
\mathbb{E}[g(Y_{i0})|D_i=1,X] &= \mathbb{E}[g(Y_{i0}(0,1))|D_i=1,X] \\
&+ \mathbb{E}[A_i(g(Y_{i0}(1,1))-g(Y_{i0}(0,1)))|D_i=1,X] \\
&= \mathbb{E}[g(Y_{i0}(0,1))|D_i=1,X]+\mathbb{P}[A_i=1|D_i=1,X]\tau_g(x) \\
&\leq \mathbb{E}[g(Y_{i0}(0))|D_i=1,X]+\pi(x)\tau_g(x)
\end{aligned}$$

I further apply the inequality that $0\leq\tau_g(x)\leq\mu_g(x)$, and call $m_g(x)=\mathbb{E}[g(Y_{i1})-g(Y_{i0})|D_i=1,X]-\mathbb{E}[g(Y_{i1})-g(Y_{i0})|D_i=0,X]$. I get

$$\tau_g(x)\leq\mu_g(x)\leq m_g(x)+\pi(x)\tau_g(x)$$

, and I can conclude that

$$0\leq\tau_g(x)\leq\frac{m_g(x)}{1-\pi(x)} \qquad \tau_g(x)\leq\mu_g(x)\leq\frac{m_g(x)}{1-\pi(x)}$$

I notice that $\mu_g(x) = \mathbb{P}[A_i = 1|D_i = 1, X]\tau_g(x) + m_g(x) \geq m_g(x)$ and can conclude that $\mu_g(x) \in \left[m_g(x), \frac{m_g(x)}{1-\pi(x)}\right]$ or I can write it in the way

$$\mathbb{E}[\mu_g - m_g|X] \geq 0 \qquad \mathbb{E}\left[\mu_g - \frac{m_g}{1-\pi(x)}\middle|X\right] \leq 0$$

To avoid calculating so many conditional expectations, I follow Abadie (2005) and write

$$m_g(x) = \mathbb{E}[\rho_0(g(Y_1) - g(Y_0))|X = x] \qquad \rho_0 = \frac{D_i - \mathbb{P}[D_i = 1|X]}{\mathbb{P}[D_i = 1|X](1 - \mathbb{P}[D_i = 1|X])}$$

. □

Multiple Periods. In applied work, the standard two periods difference-in-differences model is not the majority and people would like to incorporate data from multiple periods to support their conclusions. It is crucial to consider the case with multiple periods besides the two-period model to accommodate anticipation in longitudinal data. I will follow the framework of Sun and Abraham (2020) and consider anticipation in a staggered adoption case with multiple periods and possibly different treatment times. Sun and Abraham (2020) analyzes this framework to address the effect of treatment effect heterogeneity in two way fixed effects model and propose a difference-in-differences form estimator. My focus is on how to incorporate anticipation in a multiple-periods model and if homogeneity assumptions are imposed and researchers use an alternative approach, a similar logic can be applied to incorporate anticipation.

Suppose the potential outcome for unit i at period t is represented by $Y_{it}(a, e)$. $i \in \{1, 2, \cdots, n\}$ indexes the unit and $e \in supp(E_i) = \{1, 2, \cdots, T, \infty\}$ denotes the date when this unit gets the first treatment and E_i is one realization of e. In a setting with staggered adoption, one unit will always get treated after the first treatment. A binary random variable A_{it} that takes value $a \in \{0, 1\}$ represents the unobservable anticipation status for unit i at period t. $e = \infty$ implies this unit has never been treated and thus belongs to the control group. I am still focusing on the pre-treatment

anticipatory behavior within the treated group which means I am only distinguishing $Y_{it}(0,e)$ and $Y_{it}(1,e)$ for $t < e$ and $e \neq \infty$. In order to incorporate anticipation in the multiple periods model, the assumptions in the two-period difference-in-differences model need some modification.

Assumption A.1. *The potential outcomes* $\{Y_{it}(a,e), E_i, A_{it}\}_{t=1}^{T}$ *are independently and identically distributed across* i *for* $(a,e) \in \{0,1\} \times \{1,2,\ldots,T,\infty\}$.

Assumption A.2. $Y_{it}(0,e) = Y_{it}(0,e') \quad for \quad t < \min\{e,e'\}$

Assumption A.1 restricts the sampling process by imposing i.i.d. restrictions and assumption A.2 restricts the anticipatory behavior to be the only reason for the future to affect the present, both in the same way as in the two-period difference-in-differences model.

Under this setup, the parameter of interest I focus on is

$$\mu_g(e,t) = \mathbb{E}[g(Y_{it}(0,e)) - g(Y_{it}(0,\infty))|E_i = e]$$

with similar restrictions on function $g(.)$. Define the corresponding anticipatory effect for anticipators as

$$\tau_g(e,t) = \mathbb{E}[g(Y_{it}(1,e)) - g(Y_{it}(0,e))|E_i = e, A_i = 1]$$

and write $\mu_g(e,t)$ as $\mu(e,t)$ and $\tau_g(e,t)$ as $\tau(e,t)$ if $g(.)$ is the identity function.

Assumption A.3. *For all* $t_1 \neq t_2$*, we have*

$$\mathbb{E}[g(Y_{it_1}(0,\infty)) - g(Y_{it_2}(0,\infty))|E_i = e] = \mathbb{E}[g(Y_{it_1}(0,\infty)) - g(Y_{it_2}(0,\infty))|E_i = \infty]$$

for all e.

Parallel trends assumption is imposed and it implies potential outcomes without anticipation and treatment will change in the same way among different groups who get treated at different times.

Following Sun and Abraham (2020), I build a difference-in-differences form variable to help us analyze the parameter of interest while considering anticipation. Unlike the two periods model, I do not have a natural period 0 to compare with and need to pick one by myself. Consider the following term for $t \geq e$ and $s < e$,

$$m_g(e, s, t) = \mathbb{E}[g(Y_{it}) - g(Y_{is})|E_i = e] - \mathbb{E}[g(Y_{it}) - g(Y_{is})|E_i = \infty].$$

For simplicity I define $\mathbb{P}[A_{is} = 1|E_i = e] = h(e, s)$ which denotes the probability to anticipate in period s before the treatment happens at period e. t is a post-treatment period that I am interested in, e is the period when unit i receives the treatment and s is a chosen pre-treatment period that helps us build the difference-in-differences estimator as introduced in Sun and Abraham (2020). I will discuss the choice of period s later. Impose assumptions as in the two-period difference-in-differences model to get the bounds for the treatment effect.

Assumption A.4. $0 \leq h(e, s) \leq \pi(e, s)$.

Based on the discussion in the two-period difference-in-differences model, I know that the source of distortion is those who anticipate and react to it before the treatment. In the multiple periods model, it is important to think about the probability for those who will be treated at period e to anticipate at period s chosen as the benchmark. Further, magnitude restrictions on anticipatory effect and treatment effect at different periods are also imposed.

Assumption A.5. $|\tau_g(e, s)| \leq |\mu_g(e, t)|$.

In the Assumption A.4, I can choose $\pi(e, s)$ not only based on the treatment period, but also based on the benchmark period s I pick. One common strategy in empirical work to deal with anticipation is to argue that anticipation only exists within a certain time length before the treatment and if people have rich enough data they can always drop data when people might anticipate. This is equivalent to saying pick a far enough period s and pick $\pi(e, s) = 0$ for that period in our setup. However, it is not always reasonable to drop a subset of data and claim anticipation disappears after this. When I need to consider the bound $\pi(e, s)$, one analogy to

the choice of $\mathbb{P}[D_i = 1]$ in the two-period model that measures the intensity of treatment is given by the proportion of the group that receives treatment no later than period e, $\mathbb{P}[E_i \leq e]$, which also captures the idea that people might anticipate from information produced by prior implementations of treatment besides their own future treatments. On the other hand, researchers may want to take the time gap between the benchmark period s and the treatment period e into consideration to capture the idea that the further from the treatment, the more difficult for people to anticipate and multiply a time discount factor, for example, δ^{e-s}, for a known discounting factor $\delta \in (0, 1)$. Thus one possible choice of $\pi(e, s) = \delta^{e-s}\mathbb{P}[E_i \leq e]$. Researchers can pick different $\pi(e, s)$ based on their situations and empirical backgrounds.

The choice of s also affects the validity of Assumption A.5 as the relative time gap for the period t and s to e might differ and possibly affect the strength of this magnitude assumption. Although it seems fascinating to pick an s that is really far from the treatment period to reduce the anticipation probability and restrict the anticipatory effect magnitude, it is not cost-free. Sometimes a long period of pre-treatment data is not available. Even when it is possible to do this, picking a benchmark period s that is far from the treatment period requires the parallel trends assumption to sustain in a relatively long period and increases the risk of violation.

Based on the assumptions above, I can get the following result.

Theorem A.1. *Under Assumptions A.1-A.5 and a chosen s satisfying $s < e$, the parameter of interest $\mu_g(e, t)$ is partially identified via a closed interval in the following form, $\mu_g(e, t) \in m_g(e, s, t)\times$*

$$\left[\min\left\{1, \frac{1}{1 - sgn(\tau_g(e, s)\mu_g(e, t))\pi(e, s)}\right\}, \max\left\{1, \frac{1}{1 - sgn(\tau_g(e, s)\mu_g(e, t))\pi(e, s)}\right\}\right]$$

with $m_g(e, s, t) = \mathbb{E}[g(Y_{it}) - g(Y_{is})|E_i = e] - \mathbb{E}[g(Y_{it}) - g(Y_{is})|E_i = \infty]$.

In the multiple periods model, I end up with an expression similar to the two periods model with one side being the difference-in-differences form estimator proposed by Sun and Abraham (2020) and the other side enlarging or reducing by a specific

ratio depending on the signs of the treatment and anticipatory effect and the bound for anticipation probability. Researchers can still implement a sensitivity check by changing the choice of $\pi(e,s)$ and period s to examine whether the conclusion is robust.

The proof strategy carries from two periods model and I am still taking the case where $0 \leq \tau(e,s) \leq \mu(e,t)$ as an example. For the difference-in-differences estimand $m_g(e,s,t)$

$$
\begin{aligned}
&\mathbb{E}[g(Y_{it}) - g(Y_{is})|E_i = e] - \mathbb{E}[g(Y_{it}) - g(Y_{is})|E_i = \infty] \\
&= \mathbb{E}[g(Y_{it})|E_i = e] - \mathbb{E}[g(Y_{is})|E_i = e] - \mathbb{E}[g(Y_{it}(0,\infty)) - g(Y_{is}(0,\infty))|E_i = \infty] \\
&= \mathbb{E}[g(Y_{it}(0,e)) + A_{it}(g(Y_{it}(1,e)) - g(Y_{it}(0,e)))|E_i = e] \\
&- \mathbb{E}[g(Y_{is}(0,e)) + A_{is}(g(Y_{is}(1,e)) - g(Y_{is}(0,e)))|E_i = e] \\
&- \mathbb{E}[g(Y_{it}(0,\infty)) - g(Y_{is}(0,\infty))|E_i = \infty] \\
&= \mathbb{E}[g(Y_{it}(0,e))|E_i = e] + h(e,t)\tau(e,t) - \mathbb{E}[g(Y_{is}(0,e))|E_i = e] - h(e,s)\tau(e,s) \\
&- \mathbb{E}[g(Y_{it}(0,\infty)) - g(Y_{is}(0,\infty))|E_i = e] \\
&= \mathbb{E}[g(Y_{it}(0,e)) - g(Y_{it}(0,\infty))|E_i = e] - h(e,s)\tau(e,s) = \mu_g(e,t) - h(e,s)\tau_g(e,s)
\end{aligned}
$$

From the equation above, I can introduce Assumption A.4 and A.5, and then get

$$m_g(e,s,t) \leq \mu_g(e,t) \leq \frac{m_g(e,s,t)}{1 - \pi(e,s)}$$

□

Change-In-Changes Setup. As an alternative approach to the difference-in-differences model, Athey and Imbens (2006) proposes a change-in-changes model that does not depend on the scale of dependent variables and recovers the entire counterfactual distribution of effects of the treatment on the treatment group. The change-in-changes model also incorporates nonlinear potential outcomes. The key identifying assumption is a time-invariant distribution of the unobservable variable across different groups. In this section, I generalize the change-in-changes model to incorporate anticipation.

Following the two-period difference-in-differences setup, suppose there are n units $i \in \{1, \ldots, n\}$ with two periods $t \in \{0, 1\}$. Each unit is assigned a binary treatment D_i in the second period and has an unobservable binary anticipation status A_i in the first period. The sampling process and potential outcomes have the same restriction as before. However, to incorporate nonlinear outcomes and not use the parallel trends assumption, I follow the assumptions introduced in Athey and Imbens (2006).

Assumption A.6. *The potential outcome of a unit in the absence of both anticipation and treatment for unit i at period t is affected by an unobservable random variable U_i that represents the unit i's characteristic and satisfy*

$$Y_{it}(0,0) = \varphi(U_i, t) \qquad \textit{with} \quad \varphi(u,t) \quad \textit{strictly increasing in } u \textit{ for} \quad t \in \{0,1\}.$$

Assumption A.7. $U \perp\!\!\!\perp t|D$ *and* $\mathbb{U}_1 \subseteq \mathbb{U}_0$ *where* $\mathbb{U}_t$ *represents the support of random variable* U *in period* t.

Assumption A.6 and A.7 build the change-in-changes model. Assumption A.6 requires that the potential outcomes can be captured in a single unobservable random variable U and a higher value of random variable leads to a strictly higher potential outcome. Assumption A.7 restricts that the population does not change over time as the random variable that determines the potential outcome is independent of time period conditional on the group it belongs, which is an analogy to the parallel trends assumption in difference-in-differences model. Based on this assumption, the trend in one group can be used to recover the unobservable potential trend of the other group and comparison is possible. For this nonlinear setup where one can recover the distribution information of the potential outcomes, the parameter of interest I focus on is the quantile treatment effect for the treated group at a given quantile q

$$\mu(q) = F^{-1}_{Y^1_{i1}(0,1)}(q) - F^{-1}_{Y^1_{i1}(0,0)}(q) = F^{-1}_{Y^1_{i1}(1)}(q) - F^{-1}_{Y^1_{i1}(0)}(q)$$

F_Y represents the cumulative distribution function for the random variable Y and $F_Y^{-1}(q)$ represents the q-th quantile of the random variable Y. The switch from

two indexes to one index is based on potential outcome restrictions I impose as the potential outcomes in the second period should not depend on the anticipation status. For notation simplicity, I use Y_{it}^{d} to represent the distribution of the random variable $Y_{it}|D_i = d$, and $Y_{it}^{d_2}(d_1)$ for the random variable $Y_{it}(d_1)|D_i = d_2$, and $Y_{it}^{d_2}(a, d_1)$ for the random variable $Y_{it}(a, d_1)|D_i = d_2$. Under this setup, Athey and Imbens (2006) proposes the following identification result.

Lemma A.1 (Athey and Imbens (2006) Theorem 3.1)**.** *Suppose Assumptions II.2.1, II.2.2, A.6 and A.7 hold. The distribution of* $Y_{i1}^{1}(0)$ *can be written as*

$$F_{Y_{i1}^{1}(0)}(y) = F_{Y_{i0}^{1}(0,0)}\left(F_{Y_{i0}^{0}}^{-1}\left(F_{Y_{i1}^{0}}(y)\right)\right).$$

The potential outcome $Y_{i0}^{1}(0,0)$ should have the same distribution as Y_{i0}^{1} without anticipation, and thus the quantile of $Y_{i1}^{1}(0)$ and $\mu(q)$ are identified. However, it is known from the previous discussion that the existence of anticipation causes pre-treatment distortions, and the outcomes observed are no longer good measures of potential outcomes if there is no treatment and anticipation. Following the logic in the previous section, I need some extra assumptions to help bound the quantile treatment effect:

$$\mu(q) = F_{Y_{i1}^{1}(1)}^{-1}(q) - F_{Y_{i1}^{1}(0)}^{-1}(q) = F_{Y_{i1}^{1}(1)}^{-1}(q) - F_{Y_{i1}^{0}}^{-1}\left(F_{Y_{i0}^{0}}\left(F_{Y_{i0}^{1}(0,0)}^{-1}(q)\right)\right)$$

The first term can be estimated from the treated group, while the second term $F_{Y_{i0}^{1}(0,0)}^{-1}(q)$ is not identified from the data. Here I introduce the corresponding quantile anticipatory effect

$$\tau(q) = F_{Y_{i0}^{1}(1,1)}^{-1}(q) - F_{Y_{i0}^{1}(0,0)}^{-1}(q)$$

and use a similar strategy.

Assumption A.8. $\mathbb{P}[A_i = 1|D_i = 1] \leq \pi$.

Assumption A.9. $|\tau(q)| \leq |\mu(q)|$

Assumption A.8 is identical to the assumption I impose in the two periods difference-in-differences model. Assumption A.9 is also similar to what has been

imposed and the only difference is now I am restricting the magnitude relationship between the quantile anticipatory effect and quantile treatment effect. Based on the assumptions above, I can build bounds for the quantile treatment effect in the change-in-changes model while incorporating anticipation.

Theorem A.2. *Under Assumptions II.2.1, II.2.2 and Assumptions A.6-A.9, the parameter of interest, $\mu(q)$, is partially identified via a closed interval. For the sake of notation simplicity, let us define*

$$m(q) = F^{-1}_{Y^1_{i1}(1)}(q) - F^{-1}_{Y^0_{i1}}\left(F_{Y^0_{i0}}\left(F^{-1}_{Y^1_{i0}}(q)\right)\right)$$

$\phi_u(q)$*is the closest to zero solution for* $F^{-1}_{Y^1_{i1}(1)}(q) - F^{-1}_{Y^0_{i1}}\left(F_{Y^0_{i0}}\left(F^{-1}_{Y^1_{i0}}(q) - x\right)\right) - x = 0$

$\phi_l(q)$*is the closest to zero solution for* $F^{-1}_{Y^1_{i1}(1)}(q) - F^{-1}_{Y^0_{i1}}\left(F_{Y^0_{i0}}\left(F^{-1}_{Y^1_{i0}}(q) + x\right)\right) - x = 0$

With the restriction that $\phi_u(q)$ and $\phi_l(q)$ have the same signs as $\mu(q)$. Further define

$$\tilde{\phi}_u(q) = \begin{cases} F^{-1}_{Y^1_{i1}(1)}(q) - F^{-1}_{Y^0_{i1}}\left(F_{Y^0_{i0}}\left(F^{-1}_{Y^1_{i0}}(q-\pi)\right)\right) & q > \pi \\ +\infty & q \leq \pi \end{cases}$$

$$\tilde{\phi}_l(q) = \begin{cases} F^{-1}_{Y^1_{i1}(1)}(q) - F^{-1}_{Y^0_{i1}}\left(F_{Y^0_{i0}}\left(F^{-1}_{Y^1_{i0}}(q+\pi)\right)\right) & q < 1-\pi \\ -\infty & q \geq 1-\pi \end{cases}$$

Then we have:

$$0 \leq \tau(q) \leq \mu(q) \qquad \mu(q) \in [m(q), \min\{\phi_u(q), \tilde{\phi}_u(q)\}]$$

$$\tau(q) \leq 0 \leq \mu(q) \qquad \mu(q) \in [\max\{\phi_l(q), \tilde{\phi}_l(q)\}, m(q)]$$

$$\mu(q) \leq 0 \leq \tau(q) \qquad \mu(q) \in [m(q), \min\{\phi_l(q), \tilde{\phi}_u(q)\}]$$

$$\mu(q) \leq \tau(q) \leq 0 \qquad \mu(q) \in [\max\{\phi_u(q), \tilde{\phi}_l(q)\}, m(q)]$$

Proof of Theorem A.2. Like before, I am taking the case $0 \leq \tau(q) \leq \mu(q)$ for example. For the lower bound, I know that $\tau(q) \geq 0$ so I can conclude that

$F^{-1}_{Y^1_{i0}(0,0)}(q) \leq F^{-1}_{Y^1_{i0}}(q)$ and will have

$$\mu(q) \geq F^{-1}_{Y^1_{i1}(1)}(q) - F^{-1}_{Y^0_{i1}}\left(F_{Y^0_{i0}}\left(F^{-1}_{Y^1_{i0}}(q)\right)\right)$$

which is identifiable from observable variables. On the other hand, I can build the upper bound for $\mu(q)$ through that $F^{-1}_{Y^1_{i0}}(q) \leq F^{-1}_{Y^1_{i0}(0,0)}(q) + \tau(q) \leq F^{-1}_{Y^1_{i0}(0,0)}(q) + \mu(q)$ using magnitude restriction. Then the upper bound for $\mu(q)$, represented by $\phi_u(q)$ can be solved by the equation

$$\phi_u(q) = F^{-1}_{Y^1_{i1}(1)}(q) - F^{-1}_{Y^0_{i1}}(F_{Y^0_{i0}}(F^{-1}_{Y^1_{i0}}(q) - \phi_u(q)))$$

when $\phi_u(q) = 0$, left is smaller or equal than the right hand side. Both side are increasing functions so the minimum nonnegative solution should be $\phi_u(q)$. It is worth mentioning that $\phi_u(q)$ is not guaranteed to exist in this approach and it depends on the distribution of all these random variables. For this approach, I do not use the information from Assumption A.9. If I would like to introduce that assumption, I can make an improvement on $\phi_u(q)$ for some circumstances.

If $q \leq \pi$, then there is not much thing I can do to improve as the worst case is that all the people who anticipate lies in the lowest q percentage and you gain no information about the original distribution without anticipation. If $q > \pi$, then at most π of the people exceed the original q-th quantile of $Y^1_{i0}(0,0)$ after the anticipation so I should have that $F^{-1}_{Y^1_{i0}}(q-\pi) \leq F^{-1}_{Y^1_{i0}(0,0)}(q)$ and then I have

$$\mu(q) \leq F^{-1}_{Y^1_{i1}(1)}(q) - F^{-1}_{Y^0_{i1}}\left(F_{Y^0_{i0}}\left(F^{-1}_{Y^1_{i0}}(q-\pi)\right)\right) = \tilde{\phi}_u(q)$$

For the case $0 \leq \tau(q) \leq \mu(q)$, I will have the lower and upper bound for $\mu(q)$ as

$$\theta_l(q) = F^{-1}_{Y^1_{i1}(1)}(q) - F^{-1}_{Y^0_{i1}}\left(F_{Y^0_{i0}}\left(F^{-1}_{Y^1_{i0}}(q)\right)\right) \qquad \theta_u(q) = \min\{\phi_u(q), \tilde{\phi}_u(q)\}$$

□

From the form of the intervals, one can find that the way to build bounds for the

quantile treatment effect is different from what I have done in the linear outcome case. The key difference is that I need both of these two assumptions to build the identified set in the difference-in-differences model while in the change-in-changes setup, either restriction on anticipation probability or magnitude anticipation is possible to provide an identified set. $\phi_u(q)$ and $\phi_l(q)$ are possible bounds based on the magnitude restrictions following the idea that to the maximum extent, the pre-treatment potential outcomes without anticipation and treatment deviate from the observed pre-treatment outcomes up to a magnitude that is equal to the treatment effect. If one can find a solution to the formula, then the solution will be the corresponding bounds for the treatment effect. $\tilde{\phi}_u(q)$ and $\tilde{\phi}_l(q)$ are bounds obtained from the restrictions on anticipation probability. The intuition for the bounds is that if one is interested in the performance of q-th quantile treatment effect and the proportion of those anticipate is up to π, then at least $(q-\pi)$ of the observation is not affected by anticipation. For example, if the anticipation effect is positive, then the q-th quantile of potential outcomes without anticipation and treatment should be no less than $(q-\pi)$-th quantile of observed outcomes as at least this part of units don't anticipate and there is no distortion. Similar logic can be used to analyze other cases. Although it seems that relying on fewer assumptions represents an advantage over the difference-in-differences model and having two approaches can provide a tighter bound if researchers are willing to impose both assumptions, this approach has its own problem. Either the solution to the formula or the effectiveness of the bound achieved from the probability restriction is not guaranteed. The former one depends on the specific distribution properties of the potential outcomes and support of random variables while the latter one depends on the relationship between q one is interested in and π one picks. For example if $q \leq \pi$ a negative quantile does not provide any useful information and this bound is useless. For the purpose of covering more cases, I suggest imposing two assumptions and picking the tighter one as the identified set while applying this method.

□

Empirical Related Tables. This section provides extra empirical results for the em-

pirical application using data from subject specific teachers. □

Table A.1: Effects of the Early Retirement Incentive Program on Scores

	Original Results		With Anticipation	
	Math	Reading	Math	Reading
All Grade	0.013	0.013	[0.008,0.013]	[0.008,0.013]
	(0.008)	(0.007)		
	[-0.002,0.028]	[-0.001,0.027]	[-0.005,0.026]	[-0.004,0.025]
Grade 3	0.01	-0.003	[0.007,0.010]	[-0.007,-0.003]
	(0.013)	(0.01)		
	[-0.015,0.035]	[-0.024,0.017]	[-0.016,0.033]	[-0.05,0.04]
Grade 6	0.004	0.016	[0.002,0.004]	[0.010,0.016]
	(0.01)	(0.009)		
	[-0.016,0.024]	[-0.002,0.035]	[-0.017,0.023]	[-0.006,0.033]
Grade 8	0.032	0.03	[0.020,0.032]	[0.019,0.03]
	(0.018)	(0.017)		
	[-0.004,0.067]	[-0.003,0.063]	[-0.011,0.063]	[-0.01,0.059]

Notes: This table contains results using data from subject specific teachers. Each column presents results from a separate regression. Teachers who teach multiple grades are included in each grade. Teachers who teach in self-contained classrooms are assumed to teach both math and English. I list identified sets in the first row and 95% level confidence sets in the third row for each result with anticipation. For comparison purposes, I also provide estimators, standard errors and 95% confidence intervals for results from Fitzpatrick and Lovenheim (2014). Standard errors are displayed with parentheses.

Estimation and Inference. Here I will calculate these estimators for inference explicitly. To be consistent with the setup in empirical application, I work on the case where $\tau_g \leq 0 \leq \mu_g$ as an example and the other cases can be analyzed similarly. When I have these assumptions I know that $\mu_g \in [\mu_{g,l}, \mu_{g,u}]$ where $\mu_{g,l} = \frac{m_g}{1+\pi}$ while $\theta_u = m_g$. Corresponding estimators for lower and upper bound of the interval will

be

$$\begin{aligned}\hat{\mu}_{g,u} &= \frac{1}{n_1}\sum_{i=1}^{n}[g(Y_{i1}) - g(Y_{i0})]D_i - \frac{1}{n_0}\sum_{i=1}^{n}[g(Y_{i1}) - g(Y_{i0})](1 - D_i) \\ \hat{\mu}_{g,l} &= \frac{\hat{\mu}_{g,u}}{1+\hat{\pi}} \quad n_1 = \sum_{i=1}^{n} D_i \qquad n_0 = n - n_1 \quad \hat{\pi} \text{ is a consistent estimator for } \pi\end{aligned}$$

For simplicity let me define $\hat{\delta}_{dt} = \frac{1}{n_d}\sum_{k=1}^{n} g(Y_{kt})\mathbb{I}[D_k = d]$ for $(d,t) \in \{0,1\}^2$. I use d to index the group and t for time period here. I can also define

$$\hat{\sigma}^2_{dt} = \frac{1}{n_d - 1}\sum_{k=1}^{n}[g(Y_{kt}) - \hat{\delta}_{dt}]^2\mathbb{I}[D_k = d]$$

$$\hat{\text{cov}}_d = \frac{1}{n_d - 1}\sum_{k=1}^{n}[g(Y_{k1}) - \hat{\delta}_{d1}][g(Y_{k0}) - \hat{\delta}_{d0}]\mathbb{I}[D_k = d]$$

which measure the variance for group d at period t and covariance between two time periods for group d. For well behaved function $g(.)$, which guarantee the asymptotic normality of average, I will have

$$\hat{\sigma}^2_u = \frac{\hat{\sigma}^2_{11} + \hat{\sigma}^2_{10} - 2\hat{\text{cov}}_1}{\hat{p}} + \frac{\hat{\sigma}^2_{01} + \hat{\sigma}^2_{00} - 2\hat{\text{cov}}_0}{1 - \hat{p}}$$

$$\hat{\sigma}^2_l = \frac{\hat{\sigma}^2_{11} + \hat{\sigma}^2_{10} - 2\hat{\text{cov}}_1}{\hat{p}(1+\hat{\pi})^2} + \frac{\hat{\sigma}^2_{01} + \hat{\sigma}^2_{00} - 2\hat{\text{cov}}_0}{(1 - \hat{p})(1+\hat{\pi})^2}$$

with $\hat{p} = \frac{n_1}{n}$.

For the validity of the inference procedure, when Assumption (i) and (ii) are satisfied and the upper and lower bound estimators are ordered by construction, the procedure of Imbens and Manski (2004) is valid by Stoye (2009). The only thing need to be addressed is that now the variance of the lower and upper bounds might change so to guarantee the effectiveness of the confidence set, one needs to choose the larger variance for both bounds.

Proof of Theorem II.2. To keep consistency with the estimation procedure, concen-

trate on the case

$$\mu_g \in \left[\min\left\{m_g, \frac{m_g}{1+\pi}\right\}, \max\left\{m_g, \frac{m_g}{1+\pi}\right\}\right] \qquad \hat{\mu}_{g,u} = \hat{m}_g \qquad \hat{\mu}_{g,l} = \frac{\hat{\mu}_{g,u}}{1+\hat{\pi}}$$

If $m_g > 0$ thus $\mu_g > 0$, then based on Assumption (i) and (ii) plus the ordered upper and lower bounds estimators, the proposed interval is effective automatically following Imbens and Manski (2004) and Stoye (2009). However, when $m_g < 0$ thus $\mu_g < 0$ but I get $\hat{m}_g > 0$, then although $\hat{\mu}_{g,u}$ is still larger, now it is the estimator for the lower bound. Intuitively, when this situation happens, one can find that my inference strategy uses a smaller estimator to construct the lower bound and a larger estimator for the upper bound. At the same time, I am using the larger standard error to calculate on both sides even though the upper and lower bound estimator may change so this method should still work well. For notation simplicity, use λ to represent $\mathbb{P}[A_i = 1|D_i = 1]$ and I know $\lambda \in [0, \pi]$. Define $\sigma = \max\{\sigma_l, \sigma_u\}$, in this setup, σ is the standard deviation of m_g.

$$\begin{aligned} & \mathbb{P} \quad \left(\frac{\hat{m}_g}{1+\hat{\pi}} - C_n\frac{\hat{\sigma}}{\sqrt{n}} \leq \frac{m_g}{1+\lambda} \leq \hat{m}_g + C_n\frac{\hat{\sigma}}{\sqrt{n}}\right) \\ = & \mathbb{P} \quad \left(\frac{\sqrt{n}\frac{\hat{m}_g(\lambda-\hat{\pi})}{1+\hat{\pi}} - C_n\hat{\sigma}(1+\lambda)}{\sigma} \leq \sqrt{n}\frac{m_g - \hat{m}_g}{\sigma} \leq \frac{\sqrt{n}\lambda\hat{m}_g + C_n\hat{\sigma}(1+\lambda)}{\sigma}\right) \end{aligned}$$

For $\varepsilon > 0$, there exists N_0 such that $N > N_0$ I have $|\frac{\hat{\sigma}-\sigma}{\sigma}| < \varepsilon$ and thus $\varepsilon > 1 - \frac{\hat{\sigma}}{\sigma}$. Then the probability has a lower bound

$$\mathbb{P}\left(\frac{\frac{\sqrt{n}\hat{m}_g(\lambda-\hat{\pi})}{1+\hat{\pi}} - C_n\sigma(1+\lambda)(1-\varepsilon)}{\sigma} \leq \sqrt{n}\frac{m_g - \hat{m}_g}{\sigma} \leq \frac{\sqrt{n}\lambda\hat{m}_g + C_n\sigma(1+\lambda)(1-\varepsilon)}{\sigma}\right)$$

Use Φ to represent the cumulative value of standard normal distribution and ϕ to represent the p.d.f for standard normal here. By Berry-Essen central limit theorem,

this term is arbitrarily close to

$$\begin{aligned}
& \Phi \left(\frac{\sqrt{n}\lambda \hat{m}_g}{\sigma} + C_n(1+\lambda)(1-\varepsilon) \right) - \Phi \left(\frac{\sqrt{n}\frac{\hat{m}_g(\lambda-\hat{\pi})}{1+\hat{\pi}}}{\sigma} - C_n(1+\lambda)(1-\varepsilon) \right) \\
= & \Phi \left(\frac{\sqrt{n}\lambda \hat{m}_g}{\sigma} + C_n(1+\lambda) \right) - \Phi \left(\frac{\sqrt{n}\frac{\hat{m}_g(\lambda-\hat{\pi})}{1+\hat{\pi}}}{\sigma} - C_n(1+\lambda) \right) + 2(1+\lambda)\varepsilon C_n \phi(\omega)
\end{aligned}$$

for some ω. C_n is bounded and ε can be arbitrarily small so the last term can be ignored. Using similar logic, the left term can be written as

$$\begin{aligned}
\geq & \Phi \left(\frac{\sqrt{n}\lambda \hat{m}_g}{\hat{\sigma}} + C_n(1+\lambda) \right) - \Phi \left(-\frac{\sqrt{n}\frac{\hat{m}_g(\hat{\pi}-\lambda)}{1+\hat{\pi}}}{\hat{\sigma}} - C_n(1+\lambda) \right) \\
- & C_0 \varepsilon \lambda \left(\sqrt{n}\frac{\hat{m}_g - m_g}{\hat{\sigma}} + \sqrt{n}\frac{m_g}{\hat{\sigma}} \right) \phi(\omega')
\end{aligned}$$

for some other ω' and constant number C_0. The first term within bracket is normally distributed and the second term is negative and I can take ε arbitrarily small. Recall that $\lambda \in [0, \pi]$ and the smallest value is taken at $\lambda = 0$. I have at last

$$\begin{aligned}
& \mathbb{P} \left(\frac{\hat{m}_g}{1+\hat{\pi}} - C_n \frac{\hat{\sigma}}{\sqrt{n}} \leq \frac{m_g}{1+\lambda} \leq \hat{m}_g + C_n \frac{\hat{\sigma}}{\sqrt{n}} \right) \\
\geq & \Phi \left(C_n \right) - \Phi \left(-\frac{\sqrt{n}(\hat{\mu}_{g,u} - \hat{\mu}_{g,l})}{\hat{\sigma}} - C_n \right) = \alpha
\end{aligned}$$

□

Proof of Corollary II.1. Take the case $\tau_g \leq 0 \leq \mu_g$ for example. The new confidence set has the form

$$\left[\frac{\hat{\mu}_{g,u}}{1+\hat{\pi}} - C_n \frac{\hat{\sigma}}{\sqrt{n}}, \hat{\mu}_{g,u} + C_n \frac{\hat{\sigma}}{\sqrt{n}} \right]$$

with

$$\Phi \left(C_n + \sqrt{n}\frac{\hat{\mu}_{g,u} - \hat{\mu}_{g,l}}{\hat{\sigma}} \right) - \Phi(-C_n) = \Phi \left(C_n + \frac{\hat{\pi}}{1+\hat{\pi}} \tilde{t} \right) - \Phi(-C_n) = \alpha.$$

Given that the right hand side of the confidence set is always positive, I only need to compare $\frac{\hat{\mu}_{g,u}}{1+\hat{\pi}} - C_n\frac{\hat{\sigma}}{\sqrt{n}}$ and 0.

$$\frac{\hat{\mu}_{g,u}}{1+\hat{\pi}} - C_n\frac{\hat{\sigma}}{\sqrt{n}} = \frac{\hat{\sigma}}{(1+\hat{\pi})\sqrt{n}}(\tilde{t} - C_n(1+\hat{\pi}))$$

I know that $\Phi\left(C_n + \frac{\hat{\pi}}{1+\hat{\pi}}\tilde{t}\right) - \Phi(-C_n)$ is increasing in both C_n and $\tilde{t}$ so the solution for C_n is a decreasing function for $\tilde{t}$ at any given π. For any specific π, the smallest $\tilde{t}$ that guarantees $\frac{\tilde{t}}{1+\pi} \geq C_n$ is the value that solves

$$\Phi\left(\tilde{t}\right) - \Phi\left(-\frac{\tilde{t}}{1+\pi}\right) = \alpha.$$

The left handside is an increasing function in $\tilde{t}$ and a decreasing function in π so the worst case happens at $\pi = 1$ and that gives you the expression $\Phi(t^*) - \Phi(-t^*/2) = \alpha$. As long as $\tilde{t} > t^*$ then $\frac{\tilde{t}}{1+\pi} > C_n$ is guaranteed and I can conclude that $0 \notin CS^{\mu}_{\alpha}$. □

□

Alternative Assumptions and Corresponding Bounds. Sometimes, it might be reasonable to propose specific assumptions other than what have already been assumed. In this part, I will try to discuss several alternative assumptions that will also give partial identification results under different circumstances.

The first alternative approach mentioned here identifies the parameter of interest through a boundary on the outcomes. The advantage is that it falls naturally to certain setups, for example, binary outcomes and setups when the outcomes are scores and does not need to make discussions based on different signs of the treatment effects and do not need to bound the magnitude of effect. It will also benefit from a certain choice of bounded $g(.)$ function. However, I may need some other assumptions to validate it.

Assumption A.10. $\mathbb{E}[g(Y_{i0}(0,0))|D_i = 1, A_i = 0] = \mathbb{E}[g(Y_{i0}(0,0))|D_i = 1, A_i = 1]$

This assumption states that for those who will get treated, their potential outcome without treatment in the first period should be the same across those who anticipate

and those who do not. This is a weaker assumption compared with independent anticipation, and I only restrict this for the treated group.

Assumption A.11. *The outcome variable in the first period is bounded, which means* $g(Y_{i0}) \in [a, b]$.

The bounded outcome assumption is not quite restrictive and fits into different situations, for example, the binary outcome case. It will naturally have a bounded outcome between 0 and 1. For example, in other cases, when the outcome is grade score or the $g(.)$ function itself is bounded, this assumption is automatically satisfied.

Theorem A.3. *Under Assumptions II.2.1-II.2.3, Assumption A.10 and A.11, the parameter of interest* μ_g *is partially identified in a closed interval. The lower bound and upper bounds of that interval* $\mu_{g,l}$, $\mu_{g,u}$ *have the following form*

$$\mu_{g,l} = \mathbb{E}\left[\frac{D_i - \mathbb{P}[D_i = 1]}{\mathbb{P}[D_i = 1](1 - \mathbb{P}[D_i = 1])} g(Y_{i1}) + \frac{1 - D_i}{1 - \mathbb{P}[D_i = 1]} g(Y_{i0}) - b\right]$$

$$\mu_{g,u} = \mathbb{E}\left[\frac{D_i - \mathbb{P}[D_i = 1]}{\mathbb{P}[D_i = 1](1 - \mathbb{P}[D_i = 1])} g(Y_{i1}) + \frac{1 - D_i}{1 - \mathbb{P}[D_i = 1]} g(Y_{i0}) - a\right]$$

The proof is relatively straightforward as I can see that for the unobserved term, I will have

$$\mathbb{E}[g(Y_{i0}(0,0))|D_i = 1] = \mathbb{E}[g(Y_{i0})|D_i = 1, A_i = 0]$$

under Assumption A.10, which can help build a bound by conditional expectation. The case with covariates is similar after making all the things conditionally and is skipped here.

One may argue that the identified set given by bounded outcome might be too loose in some circumstances as it just used the upper and lower bound of the outcome itself. If people do not like to put restrictions on the signs and magnitudes of treatment effects, I can also provide a corresponding identification result under some other assumptions. The idea is that, I may not observe the anticipation status for everybody, however, if I have a bound for the anticipation probability among the group I am interested in, I can get an bound about it by assigning anticipation treat-

ment to those with highest or lowest outcomes. As among all possible anticipation treatment status satisfying the proportion restriction, I am picking the worst and best one based on the observed outcomes so I can get corresponding upper and lower bounds.

Theorem A.4. *Under Assumption II.2.1-II.2.3, II.3.1 and Assumption A.10, I can build a closed interval for the parameter of interest μ_g with upper and lower bounds $\mu_{g,l}$ and $\mu_{g,u}$ satisfying*

$$\mu_{g,l} = \mathbb{E}\left[\frac{D_i - \mathbb{P}[D_i=1]}{\mathbb{P}[D_i=1](1-\mathbb{P}[D_i=1])} g(Y_{i1}) + \frac{1-D_i}{1-\mathbb{P}[D_i=1]} g(Y_{i0})\right]$$
$$-\mathbb{E}[g(Y_{i0})|D_i=1, g(Y_{i0}) \geq g(Y_{i0})_\eta]$$
$$\mu_{g,u} = \mathbb{E}\left[\frac{D_i - \mathbb{P}[D_i=1]}{\mathbb{P}[D_i=1](1-\mathbb{P}[D_i=1])} g(Y_{i1}) + \frac{1-D_i}{1-\mathbb{P}[D_i=1]} g(Y_{i0})\right]$$
$$-\mathbb{E}[g(Y_{i0})|D_i=1, g(Y_{i0}) \leq g(Y_{i0})_{1-\eta}]$$

$g(Y_{i0})_\alpha$ denotes the α-th quantile of $g(Y_{i0})$ conditional on $D_i = 1$ and $\eta = \pi$.

The proof for this theorem follows the idea that I can assign the anticipation treatment group to those with highest or lowest outcomes under a bounded proportion and this will provide bounds for the purpose of partial identification. For these bounds, all the items are observable and can be estimated and I do not need to put restrictions on signs and magnitudes of the treatment effect. □

APPENDIX B

Proofs for Chapter III

First we introduce some technical lemmas that are used to establish our main results. The main results are obtained by working with the representation

$$\sqrt{n}(\hat{\beta}_n - \beta) = \hat{\mathbf{\Gamma}}_n^{-1}\mathbf{S}_n,$$

where

$$\hat{\mathbf{\Gamma}}_n = \frac{1}{n}\sum_{1\leq i\leq n}\hat{\mathbf{v}}_{i,n}\hat{\mathbf{v}}_{i,n}'$$

and

$$\mathbf{S}_n = \frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}\hat{\mathbf{v}}_{i,n}u_{i,n} = \frac{1}{\sqrt{n}}\sum_{1\leq g\leq N_{\mathcal{T},n}}\hat{\mathbf{V}}_n(g)'\mathbf{u}_n(g),$$

$$\mathbf{u}_n\left(g\right) = \left(u_{t_{g,n}(1),n},\ldots,u_{t_{g,n}(\#\mathcal{T}_{g,n}),n}\right)'.$$

Strictly speaking, the displayed representation is valid only when $\lambda_{\min}(\sum_{i=1}^n \mathbf{w}_{i,n}\mathbf{w}_{i,n}') > 0$ and $\lambda_{\min}(\hat{\mathbf{\Gamma}}_n) > 0$. Both events occur with probability approaching one under our assumptions, so without loss of generality we may assume that they occur almost surely.

The first lemma can be used to bound $\hat{\mathbf{\Gamma}}_n^{-1}$.

Lemma B.1. *If Assumptions III.2.1-III.2.3 hold, then* $\hat{\mathbf{\Gamma}}_n^{-1} = O_p(1)$.

Let $\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_n(\mathcal{X}_n, \mathcal{W}_n) = \mathbb{V}[\mathbf{S}_n|\mathcal{X}_n, \mathcal{W}_n]$. The second lemma can be used to bound $\boldsymbol{\Sigma}_n^{-1}$ and to show asymptotic normality of $\mathbf{S}_n$.

Lemma B.2. *Suppose Assumptions III.2.1-III.2.3 hold and suppose that*

$$\mathcal{C}_{\mathcal{S},n}\rho_n = o(1) \qquad \textit{and} \qquad \frac{\mathcal{C}_{\mathcal{T},n}^3}{n^2}\sum_{1\leq i\leq n}||\hat{\mathbf{v}}_{i,n}||^4 = o_p(1).$$

Then $\boldsymbol{\Sigma}_n^{-1} = O_p(1)$ *and* $\boldsymbol{\Sigma}_n^{-1/2}\mathbf{S}_n \to_d \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

The third lemma can be used to approximate

$$\hat{\boldsymbol{\Sigma}}_n = \hat{\boldsymbol{\Sigma}}_n(\kappa_n) = \text{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'\kappa_n(g,h)(\hat{\mathbf{u}}_n(h)\otimes\hat{\mathbf{u}}_n(h))\right\}$$

by means of

$$\bar{\boldsymbol{\Sigma}}_n = \bar{\boldsymbol{\Sigma}}_n(\kappa_n; \mathcal{X}_n, \mathcal{W}_n) = \mathbb{E}[\tilde{\boldsymbol{\Sigma}}_n(\kappa_n)|\mathcal{X}_n, \mathcal{W}_n],$$

where

$$\tilde{\boldsymbol{\Sigma}}_n(\kappa_n) = \text{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'\kappa_n(g,h)(\tilde{\mathbf{U}}_n(h)\otimes\tilde{\mathbf{U}}_n(h))\right\},$$

$$\tilde{\mathbf{U}}_n(h) = (\tilde{U}_{t_{h,n}(1),n}, \ldots, \tilde{U}_{t_{h,n}(\#\mathcal{T}_{h,n}),n})', \qquad \tilde{U}_{i,n} = \sum_{1\leq j\leq n} M_{ij,n}U_{j,n}.$$

Lemma B.3. *Suppose Assumptions III.2.1-III.2.3 hold and suppose that*

$$\mathcal{C}_{\mathcal{T},n}^3[\mathcal{C}_{\mathcal{S},n}\rho_n + n(\varrho_n - \rho_n) + n\chi_n\varrho_n] = o(1) \qquad \textit{and} \qquad \frac{\mathcal{C}_{\mathcal{T},n}^4}{n^2}\sum_{1\leq i\leq n}||\hat{\mathbf{v}}_{i,n}||^4 = o_p(1).$$

If $\|\kappa_n\|_\infty = O_p(1)$, *then* $\hat{\boldsymbol{\Sigma}}_n = \bar{\boldsymbol{\Sigma}}_n + o_p(1)$.

The fourth lemma can be combined with the third lemma to show consistency of $\hat{\boldsymbol{\Sigma}}_n$.

Lemma B.4. *Suppose Assumptions III.2.1 and III.2.2 hold and suppose that* $\chi_n = O(1)$. *If*

$$\mathcal{C}_{\mathcal{T},n} \left|\left|\kappa_n(\mathbf{M}_n \circledast_n \mathbf{M}_n) - \mathbf{I}_{N_{\kappa,n}}\right|\right|_\infty = o_p(1),$$

then $\bar{\mathbf{\Sigma}}_n = \mathbf{\Sigma}_n + o_p(1)$.

Lemmas 3 and 4 make high-level assumptions about κ_n. The fifth lemma gives sufficient conditions for the assumptions for specific κ_n.

Lemma B.5. *Suppose Assumptions III.2.1 and III.2.2 hold and suppose that* $\chi_n = O(1)$.

(a) If $\kappa_n = \mathbf{I}_{N_{\kappa,n}}$, *then* $\|\kappa_n\|_\infty = O_p(1)$. *If also* $\mathcal{C}^2_{\mathcal{T},n}\mathcal{M}_n = o_p(1)$, *then* $\bar{\mathbf{\Sigma}}_n = \mathbf{\Sigma}_n + o_p(1)$.

(b) Suppose $\kappa_n = \kappa_n^{\texttt{BR}}$. *If*

$$\lim_{\delta \downarrow 0} \liminf_{n\to\infty} \mathbb{P}\left[\min_{1\le g\le N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{-1/2}||_\infty^{-2} > \delta\right] = 1,$$

then $\|\kappa_n\|_\infty = O_p(1)$. *If also* $\mathcal{C}^3_{\mathcal{T},n}\mathcal{M}_n = o_p(1)$, *then* $\bar{\mathbf{\Sigma}}_n = \mathbf{\Sigma}_n + o_p(1)$.

(c) Suppose $\kappa_n = \kappa_n^{\texttt{JK}}$. *If*

$$\lim_{\delta \downarrow 0} \liminf_{n\to\infty} \mathbb{P}\left[\min_{1\le g\le N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} > \delta\right] = 1,$$

then $\|\kappa_n\|_\infty = O_p(1)$ *and* $\bar{\mathbf{\Sigma}}_n \ge \mathbf{\Sigma}_n + o_p(1)$.

If $\mathcal{C}^2_{\mathcal{T},n}\mathcal{M}_n = o_p(1)$, *then* $\|\kappa_n\|_\infty = O_p(1)$ *and* $\bar{\mathbf{\Sigma}}_n = \mathbf{\Sigma}_n + o_p(1)$.

(d) Suppose $\kappa_n = \kappa_n^{\texttt{CR}}$. *If*

$$\lim_{\delta \downarrow 0} \liminf_{n\to\infty} \mathbb{P}\left[\min_{1\le g\le N_{\mathcal{T},n}} \left\{ ||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} - \sum_{1\le h\le N_{\mathcal{T},n}, h\ne g} ||\mathbf{M}_n(g,h)||_\infty^2 \right\} > \delta\right] = 1,$$

then $\|\kappa_n\|_\infty = O_p(1)$ *and* $\bar{\mathbf{\Sigma}}_n = \mathbf{\Sigma}_n + o_p(1)$.

Proof of Theorem III.1. Theorem III.1 follows from Lemmas B.1 and B.2.

□

Proof of Theorem III.2. Theorem III.2 follows from combining Theorem III.1 with Lemmas B.3, B.4, and B.5. □

Proof of Corollary III.1. The function $\mathbf{M} \mapsto ||\mathbf{M}^{-1/2}||_\infty^{-2}$ is continuous at $\mathbf{M} = \mathbf{I}$. As a consequence, if $\mathcal{C}_{\mathcal{T},n} = O(1)$ and if $\mathcal{M}_n = o_p(1)$, then

$$\min_{1 \leq g \leq N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{-1/2}||_\infty^{-2} = 1 + o_p(1),$$

implying in particular that

$$\lim_{\delta \downarrow 0} \liminf_{n \to \infty} \mathbb{P}\left[\min_{1 \leq g \leq N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{-1/2}||_\infty^{-2} > \delta \right] = 1.$$

In other words, the assumptions of Theorem III.2 (a)-(b) are satisfied under the conditions of Corollary III.1. □

Proof of Corollary III.2. If $g \neq h$, then

$$||\mathbf{M}_n(g,h)||_\infty^2 \leq \mathcal{C}_{\mathcal{T},n} ||\mathbf{M}_n(g,h)||_F^2 = \mathcal{C}_{\mathcal{T},n} \sum_{1 \leq k \leq \#\mathcal{T}_{g,n}, 1 \leq K \leq \#\mathcal{T}_{h,n}} M_{t_{g,n}(k), t_{h,n}(K), n}^2,$$

and therefore

$$\begin{aligned} \sum_{1 \leq h \leq N_{\mathcal{T},n}, h \neq g} ||\mathbf{M}_n(g,h)||_\infty^2 &\leq \mathcal{C}_{\mathcal{T},n} \sum_{1 \leq k \leq \#\mathcal{T}_{g,n}} M_{t_{g,n}(k), t_{g,n}(k), n}(1 - M_{t_{g,n}(k), t_{g,n}(k), n}) \\ &\leq \mathcal{C}_{\mathcal{T},n}^2 \max_{1 \leq i \leq n} \left\{ M_{ii,n}(1 - M_{ii,n}) \right\}, \end{aligned}$$

implying in particular that if $\mathcal{M}_n < 1/2$, then

$$\sum_{1 \leq h \leq N_{\mathcal{T},n}, h \neq g} ||\mathbf{M}_n(g,h)||_\infty^2 \leq \mathcal{C}_{\mathcal{T},n}^2 (1 - \mathcal{M}_n) \mathcal{M}_n.$$

Suppose

$$(\mathcal{C}_{\mathcal{T},n}^2 - \mathcal{C}_{\mathcal{T},n} + 2)\mathcal{M}_n + \sqrt{(\mathcal{C}_{\mathcal{T},n} - 1)(1 - \mathcal{M}_n)\mathcal{M}_n} < 1.$$

Then

$$\mathcal{C}_{\mathcal{T},n}\mathcal{M}_n < 1 \qquad \text{and} \qquad \mathcal{M}_n < 1/2,$$

so $\mathbf{M}_n \circledast_n \mathbf{M}_n$ is block diagonally dominant because

$$\begin{aligned}
& ||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} - \sum_{1\le h\le N_{\mathcal{T},n}, h\neq g} ||\mathbf{M}_n(g,h)||_\infty^2 \\
\geq\ & (1-\mathcal{M}_n)^2 + (\mathcal{C}_{\mathcal{T},n}-1)(1-\mathcal{M}_n)\mathcal{M}_n - (1-\mathcal{M}_n)\sqrt{(\mathcal{C}_{\mathcal{T},n}-1)(1-\mathcal{M}_n)\mathcal{M}_n} \\
& -\mathcal{C}_{\mathcal{T},n}^2(1-\mathcal{M}_n)\mathcal{M}_n \\
=\ & (1-\mathcal{M}_n)\left\{1 - (\mathcal{C}_{\mathcal{T},n}^2 - \mathcal{C}_{\mathcal{T},n} + 2)\mathcal{M}_n - \sqrt{(\mathcal{C}_{\mathcal{T},n}-1)(1-\mathcal{M}_n)\mathcal{M}_n}\right\} \\
>\ & 0.
\end{aligned}$$

In particular,

$$\begin{aligned}
& \lim_{\delta\downarrow 0}\liminf_{n\to\infty} \mathbb{P}\left[\min_{1\le g\le N_{\mathcal{T},n}}\left\{||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} - \sum_{1\le h\le N_{\mathcal{T},n}, h\neq g} ||\mathbf{M}_n(g,h)||_\infty^2\right\} > \delta\right] \\
\geq\ & \lim_{\delta\downarrow 0}\liminf_{n\to\infty} \mathbb{P}\left[(\mathcal{C}_{\mathcal{T},n}^2 - \mathcal{C}_{\mathcal{T},n} + 2)\mathcal{M}_n + \sqrt{(\mathcal{C}_{\mathcal{T},n}-1)(1-\mathcal{M}_n)\mathcal{M}_n} < 1-\delta\right],
\end{aligned}$$

so the assumptions of Theorem III.2 (c) are satisfied under the conditions of Corollary III.2. □

Proof of Corollary III.3. It is shown in the proof of Lemma B.5 (c) that

$$\lim_{\delta\downarrow 0}\liminf_{n\to\infty} \mathbb{P}\left[\min_{1\le g\le N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} > \delta\right] \geq \lim_{\delta\downarrow 0}\liminf_{n\to\infty} \mathbb{P}\left[\mathcal{C}_{\mathcal{T},n}\mathcal{M}_n < 1-\delta\right].$$

Therefore, the assumptions of Theorem III.2 (d) are satisfied under the conditions of Corollary III.3. □

Proof of Theorem III.3. If the design is cluster-orthogonal, then $\tilde{\mathbf{U}}_n(g) = \mathbf{M}_n(g,g)\mathbf{U}_n(g)$ and $\hat{\mathbf{V}}_n(g) = \mathbf{M}_n(g,g)\mathbf{X}_n(g)$, where $\mathbf{X}_n(g) = (\mathbf{x}_{t_{g,n}(1),n}, \ldots, \mathbf{x}_{t_{g,n}(\#\mathcal{T}_{g,n}),n})'$ and where each $\mathbf{M}_n(g,g)$ is idempotent. and therefore $\mathbf{M}_n(g,g)^{1/2} = \mathbf{M}_n(g,g), \mathbf{M}_n(g,g)^2 =$

$\mathbf{M}_n(g,g)$, and $\mathbf{M}_n(g,g)^+ = \mathbf{M}_n(g,g)$.

It follows from Lemma B.3 that

$$\hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}} = \hat{\mathbf{\Sigma}}_n(\mathbf{I}_{N_{\kappa,n}}) = \mathbb{E}[\tilde{\mathbf{\Sigma}}_n(\mathbf{I}_{N_{\kappa,n}})|\mathcal{X}_n, \mathcal{W}_n] + o_p(1),$$

where $\mathbb{E}[\tilde{\mathbf{\Sigma}}_n(\mathbf{I}_{N_{\kappa,n}})|\mathcal{X}_n, \mathcal{W}_n] = \mathbf{\Sigma}_n$ because

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_n(\mathbf{I}_{N_{\kappa,n}}) &= \mathrm{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\le g\le N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))(\tilde{\mathbf{U}}_n(g)\otimes\tilde{\mathbf{U}}_n(g))\right\}\\ &= \mathrm{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\le g\le N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))(\mathbf{U}_n(g)\otimes\mathbf{U}_n(g))\right\},\end{aligned}$$

the second equality using

$$\hat{\mathbf{V}}_n(g)'\tilde{\mathbf{U}}_n(g) = \mathbf{X}_n(g)'\mathbf{M}_n(g,g)^2\mathbf{U}_n(g) = \mathbf{X}_n(g)'\mathbf{M}_n(g,g)\mathbf{U}_n(g) = \hat{\mathbf{V}}_n(g)'\mathbf{U}_n(g).$$

Moreover,

$$\hat{\mathbf{\Sigma}}_n^{\mathtt{BR}} = \hat{\mathbf{\Sigma}}_n^{\mathtt{JK}} = \hat{\mathbf{\Sigma}}_n^{\mathtt{CR}} = \hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}},$$

where the first equality uses $\mathbf{M}_n(g,g)^{1/2} = \mathbf{M}_n(g,g)$, the second equality uses block diagonality of $\kappa_n^{\mathtt{CR}}$, and the last equality uses

$$\mathbf{M}_n(g,g)^+ = \mathbf{M}_n(g,g), \qquad \hat{\mathbf{V}}_n(g)'\mathbf{M}_n(g,g)\tilde{\mathbf{U}}_n(g) = \hat{\mathbf{V}}_n(g)'\mathbf{U}_n(g).$$

Finally, if $\mathcal{C}_{\mathcal{T},n}\mathcal{M}_n = o_p(1)$, then $\mu_n\mathbf{\Sigma}_n = \mathbf{\Sigma}_n + o_p(1)$ and therefore

$$\check{\mathbf{\Sigma}}_n^{\mathtt{LZ}} = \mu_n\hat{\mathbf{\Sigma}}_n^{\mathtt{LZ}} = \mathbf{\Sigma}_n + o_p(1).$$

□

Then we introduce the proof of technical lemmas. Throughout the proofs we simplify the notation by assuming without loss of generality that $d = 1$. In Lemma B.2 the case where $d > 1$ can be handled by means of the Cramér-Wold device and

simple bounding arguments.

Proof of Lemma B.1. It suffices to show that $\tilde{\Gamma}_n = \mathbb{E}[\tilde{\Gamma}_n|\mathcal{W}_n] + o_p(1)$ and that $\hat{\Gamma}_n - \tilde{\Gamma}_n \geq o_p(1)$.

First,

$$\tilde{\Gamma}_n = \frac{1}{n}\sum_{1\leq G\leq N_{\mathcal{S},n}} a_{GG,n} + \frac{2}{n}\sum_{1\leq G,H\leq N_{\mathcal{S},n},G<H} a_{GH,n}, \qquad a_{GH,n} = \sum_{s\in\mathcal{S}_{G,n},t\in\mathcal{S}_{H,n}} M_{st,n}V_{s,n}V_{t,n},$$

where $\sum_{1\leq G,H\leq N_{\mathcal{S},n}} \mathbb{V}[a_{GH,n}|\mathcal{W}_n] = o_p(n^2)$ because

$$\begin{aligned}\mathbb{V}[a_{GH,n}|\mathcal{W}_n] &\leq (\#\mathcal{S}_{G,n})(\#\mathcal{S}_{H,n}) \sum_{s\in\mathcal{S}_{G,n},t\in\mathcal{S}_{H,n}} M_{st,n}^2\mathbb{V}[V_{s,n}V_{t,n}|\mathcal{W}_n] \\ &\leq \mathcal{C}_{\mathcal{S},n}^2\mathcal{C}_{V,n}\sum_{s\in\mathcal{S}_{G,n},t\in\mathcal{S}_{H,n}} M_{st,n}^2,\end{aligned}$$

where $\mathcal{C}_{\mathcal{S},n} = o(\sqrt{n})$, $\mathcal{C}_{V,n} = 1 + \max_{1\leq i\leq n}\mathbb{E}[V_{i,n}^4|\mathcal{W}_n] = O_p(1)$, and

$$\sum_{1\leq G,H\leq N_{\mathcal{S},n}}\sum_{s\in\mathcal{S}_{G,n},t\in\mathcal{S}_{H,n}} M_{st,n}^2 = \sum_{1\leq i,j\leq n} M_{ij,n}^2 = \sum_{1\leq i\leq n} M_{ii,n} \leq n.$$

As a consequence,

$$\mathbb{V}\left[\frac{1}{n}\sum_{1\leq G\leq N_{\mathcal{S},n}} a_{GG,n}\middle|\mathcal{W}_n\right] = \sum_{1\leq G\leq N_{\mathcal{S},n}}\frac{\mathbb{V}[a_{GG,n}|\mathcal{W}_n]}{n^2} \leq \sum_{1\leq G,H\leq N_{\mathcal{S},n}}\frac{\mathbb{V}[a_{GH,n}|\mathcal{W}_n]}{n^2} = o_p(1)$$

and

$$\begin{aligned}\mathbb{V}\left[\frac{1}{n}\sum_{1\leq G,H\leq N_{\mathcal{S},n},G<H} a_{GH,n}\middle|\mathcal{W}_n\right] &= \frac{1}{n^2}\sum_{1\leq G,H\leq N_{\mathcal{S},n},G<H}\mathbb{V}[a_{GH,n}|\mathcal{W}_n] \\ &\leq \frac{1}{n^2}\sum_{1\leq G,H\leq N_{\mathcal{S},n}}\mathbb{V}[a_{GH,n}|\mathcal{W}_n] = o_p(1),\end{aligned}$$

implying in particular that $\tilde{\Gamma}_n = \mathbb{E}[\tilde{\Gamma}_n|\mathcal{W}_n] + o_p(1)$.

Next, defining $\tilde{Q}_{i,n} = \sum_{1\leq j\leq n} M_{ij,n} Q_{j,n}$, we have

$$\begin{aligned}\hat{\Gamma}_n - \tilde{\Gamma}_n &= \frac{1}{n}\sum_{1\leq i\leq n} \tilde{Q}_{i,n}^2 + \frac{2}{n}\sum_{1\leq i\leq n} \tilde{Q}_{i,n}\tilde{V}_{i,n} \\ &\geq \frac{2}{n}\sum_{1\leq i\leq n} \tilde{Q}_{i,n}\tilde{V}_{i,n} = \frac{2}{n}\sum_{1\leq i\leq n} \tilde{Q}_{i,n} V_{i,n} = O_p\left(\frac{\mathcal{C}_{\mathcal{S},n}\chi_n}{n}\right) = o_p(1),\end{aligned}$$

the penultimate equality using the facts that $\mathbb{E}[\tilde{Q}_{i,n}V_{i,n}|\mathcal{W}_n] = 0$ and

$$\begin{aligned}\mathbb{V}\left[\frac{1}{n}\sum_{1\leq i\leq n} \tilde{Q}_{i,n}V_{i,n}\middle|\mathcal{W}_n\right] &= \frac{1}{n^2}\sum_{1\leq G\leq N_{\mathcal{S},n}} \mathbb{V}\left[\sum_{s\in\mathcal{S}_{G,n}} \tilde{Q}_{s,n}V_{s,n}\middle|\mathcal{W}_n\right] \\ &\leq \frac{\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{V,n}}{n^2}\sum_{1\leq G\leq N_{\mathcal{S},n}}\sum_{s\in\mathcal{S}_{G,n}} \tilde{Q}_{s,n}^2 = \frac{\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{V,n}}{n}\tilde{\chi}_n,\end{aligned}$$

where

$$\tilde{\chi}_n = \frac{1}{n}\sum_{1\leq G\leq N_{\mathcal{S},n}}\sum_{s\in\mathcal{S}_{G,n}} \tilde{Q}_{s,n}^2 = \frac{1}{n}\sum_{1\leq i\leq n} \tilde{Q}_{i,n}^2 \leq \frac{1}{n}\sum_{1\leq i\leq n} Q_{i,n}^2 = O_p(\chi_n).$$

Remark B.1. • The following assumptions are used in the proof:

- $\mathcal{C}_{\mathcal{S},n} = o(\sqrt{n})$
- $\max_{1\leq i\leq n}\mathbb{E}[\|\mathbf{V}_{i,n}\|^4|\mathcal{W}_n] = O_p(1)$
- $\chi_n = O(\sqrt{n})$

□

Proof of Lemma B.2. Defining $\tilde{S}_n = S_n - \mathbb{E}[S_n|\mathcal{X}_n,\mathcal{W}_n] = \sum_{1\leq i\leq n}\hat{v}_{i,n}U_{i,n}/\sqrt{n}$ and employing the decomposition

$$S_n - \tilde{S}_n = \frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}\tilde{V}_{i,n}r_{i,n} + \frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}\tilde{Q}_{i,n}r_{i,n} + \frac{1}{\sqrt{n}}\sum_{1\leq i\leq n}\hat{v}_{i,n}(R_{i,n} - r_{i,n}),$$

we begin by showing that $S_n = \tilde{S}_n + o_p(1)$.

First, defining $\tilde{r}_{i,n} = \sum_{1\leq j\leq n} M_{ij,n} r_{j,n}$ and using $\mathbb{E}[\tilde{r}_{i,n} V_{i,n} | \mathcal{W}_n] = 0$ and

$$\mathbb{V}\left[\frac{1}{\sqrt{n}} \sum_{1\leq i\leq n} \tilde{r}_{i,n} V_{i,n} \middle| \mathcal{W}_n\right] = \frac{1}{n} \sum_{1\leq G\leq N_{\mathcal{S},n}} \mathbb{V}\left[\sum_{s\in\mathcal{S}_{G,n}} \tilde{r}_{s,n} V_{s,n} \middle| \mathcal{W}_n\right] \leq \mathcal{C}_{\mathcal{S},n} \mathcal{C}_{V,n} \tilde{\rho}_n,$$

where

$$\tilde{\rho}_n = \frac{1}{n} \sum_{1\leq i\leq n} \tilde{r}_{i,n}^2 \leq \frac{1}{n} \sum_{1\leq i\leq n} r_{i,n}^2 = O_p(\rho_n),$$

we have

$$\frac{1}{\sqrt{n}} \sum_{1\leq i\leq n} \tilde{V}_{i,n} r_{i,n} = \frac{1}{\sqrt{n}} \sum_{1\leq i\leq n} \tilde{r}_{i,n} V_{i,n} = O_p(\mathcal{C}_{\mathcal{S},n} \rho_n) = o_p(1).$$

Also, using the Cauchy-Schwarz inequality,

$$\left|\frac{1}{\sqrt{n}} \sum_{1\leq i\leq n} \tilde{Q}_{i,n} r_{i,n}\right|^2 = \left|\frac{1}{\sqrt{n}} \sum_{1\leq i\leq n} \tilde{Q}_{i,n} \tilde{r}_{i,n}\right|^2 \leq n\tilde{\chi}_n \tilde{\rho}_n = O_p(n\chi_n \rho_n) = o_p(1)$$

and

$$\left|\frac{1}{\sqrt{n}} \sum_{1\leq i\leq n} \hat{v}_{i,n}(R_{i,n} - r_{i,n})\right|^2 \leq n\hat{\Gamma}_n \left(\frac{1}{n} \sum_{1\leq i\leq n} |R_{i,n} - r_{i,n}|^2\right) = O_p[n(\varrho_n - \rho_n)] = o_p(1),$$

where the penultimate equality uses

$$\hat{\Gamma}_n = \frac{1}{n} \sum_{1\leq i\leq n} \hat{v}_{i,n}^2 \leq \frac{1}{n} \sum_{1\leq i\leq n} v_{i,n}^2 \leq \frac{2}{n} \sum_{1\leq i\leq n} Q_{i,n}^2 + \frac{2}{n} \sum_{1\leq i\leq n} V_{i,n}^2 = O_p(\chi_n + 1) = O_p(1)$$

and $\mathbb{E}[|R_{i,n} - r_{i,n}|^2] = \mathbb{E}[R_{i,n}^2] - \mathbb{E}[r_{i,n}^2]$. As a consequence, $S_n = \tilde{S}_n + o_p(1)$.

Next,

$$\begin{aligned}\Sigma_n &= \frac{1}{n}\sum_{1\le g\le N_{\mathcal{T},n}} \mathbb{V}[\hat{\mathbf{V}}_n(g)'\mathbf{U}_n(g)|\mathcal{X}_n,\mathcal{W}_n] \\ &= \frac{1}{n}\sum_{1\le g\le N_{\mathcal{T},n}} \hat{\mathbf{V}}_n(g)'\mathbb{E}[\mathbf{U}_n(g)\mathbf{U}_n(g)'|\mathcal{X}_n,\mathcal{W}_n]\hat{\mathbf{V}}_n(g) \\ &\ge \hat{\Gamma}_n \min_{1\le g\le N_{\mathcal{T},n}} \lambda_{\min}(\mathbb{E}[\mathbf{U}_n(g)\mathbf{U}_n(g)'|\mathcal{X}_n,\mathcal{W}_n]),\end{aligned}$$

so $\Sigma_n^{-1} = O_p(1)$. The proof can therefore be completed by showing that $\Sigma_n^{-1/2}\tilde{S}_n \to_d \mathcal{N}(0,1)$.

We shall do so assuming without loss of generality that $\lambda_{\min}(\Sigma_n) > 0$ (a.s.). Because

$$\Sigma_n^{-1/2}\tilde{S}_n = \frac{1}{\sqrt{n}}\sum_{1\le g\le N_{\mathcal{T},n}} \eta_n(g), \qquad \eta_n(g) = \Sigma_n^{-1/2}\hat{\mathbf{V}}_n(g)'\mathbf{U}_n(g) = \Sigma_n^{-1/2}\sum_{t\in\mathcal{T}_{g,n}} \hat{v}_{t,n}U_{t,n},$$

where, conditional on $(\mathcal{X}_n,\mathcal{W}_n)$, $\eta_n(g)$ are mean zero independent random variables with

$$\frac{1}{n}\sum_{1\le g\le N_{\mathcal{T},n}} \mathbb{V}[\eta_n(g)|\mathcal{X}_n,\mathcal{W}_n] = 1,$$

it suffices to show that the following Lyapunov condition is satisfied:

$$\frac{1}{n^2}\sum_{1\le g\le N_{\mathcal{T},n}} \mathbb{E}[\eta_n(g)^4|\mathcal{X}_n,\mathcal{W}_n] = o_p(1).$$

Now,

$$\begin{aligned}\sum_{1\leq g\leq N_{\mathcal{T},n}}\frac{\mathbb{E}[\eta_n(g)^4|\mathcal{X}_n,\mathcal{W}_n]}{n^2} &\leq \frac{\sum_{1\leq g\leq N_{\mathcal{T},n}}\mathbb{E}\left[\left(\sum_{t\in\mathcal{T}_{g,n}}\hat{v}_{t,n}U_{t,n}\right)^4\middle|\mathcal{X}_n,\mathcal{W}_n\right]}{n\lambda_{\min}(\Sigma_n)^2}\\ &\leq \frac{\mathcal{C}_{\mathcal{T},n}^3\mathcal{C}_{U,n}}{n^2\lambda_{\min}(\Sigma_n)^2}\sum_{1\leq g\leq N_{\mathcal{T},n}}\sum_{t\in\mathcal{T}_{g,n}}\hat{v}_{t,n}^4\\ &= \frac{\mathcal{C}_{U,n}}{\lambda_{\min}(\Sigma_n)^2}\left(\frac{\mathcal{C}_{\mathcal{T},n}^3}{n^2}\sum_{1\leq i\leq n}\hat{v}_{i,n}^4\right)=o_p(1),\end{aligned}$$

where $\mathcal{C}_{U,n}=1+\max_{1\leq i\leq n}\mathbb{E}[U_{i,n}^4|\mathcal{X}_n,\mathcal{W}_n]=O_p(1)$.

Remark B.2.

- The following assumptions are used in the proof:
  - $\hat{\Gamma}_n^{-1}=O_p(1)$
  - $\max_{1\leq i\leq n}\mathbb{E}[U_{i,n}^4|\mathcal{X}_n,\mathcal{W}_n]=O_p(1)$
  - $\max_{1\leq i\leq n}\mathbb{E}[\|\mathbf{V}_{i,n}\|^4|\mathcal{W}_n]=O_p(1)$
  - $\max_{1\leq g\leq N_{\mathcal{T},n}}\{1/\lambda_{\min}(\mathbb{E}[\mathbf{U}_n(g)\mathbf{U}_n(g)'|\mathcal{X}_n,\mathcal{W}_n])\}=O_p(1)$
  - $\chi_n=O(1)$
  - $\mathcal{C}_{\mathcal{S},n}\rho_n=o(1)$
  - $n(\varrho_n-\rho_n)=o(1)$
  - $n\chi_n\rho_n=o(1)$
  - $\mathcal{C}_{\mathcal{T},n}^3n^{-2}\sum_{1\leq i\leq n}||\hat{\mathbf{v}}_{i,n}||^4=o_p(1)$
- The rate of convergence of $\hat{\beta}_n$ can be slower than $\sqrt{n}$: Because

$$\begin{aligned}\Sigma_n &= \frac{1}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}}\mathbb{V}[\hat{\mathbf{V}}_n(g)'\mathbf{U}_n(g)|\mathcal{X}_n,\mathcal{W}_n]\\ &= \frac{1}{n}\sum_{1\leq i\leq N_{\mathcal{T},n}}\hat{\mathbf{V}}_n(g)'\mathbb{E}[\mathbf{U}_n(g)\mathbf{U}_n(g)'|\mathcal{X}_n,\mathcal{W}_n]\hat{\mathbf{V}}_n(g)\\ &\leq \frac{\mathcal{C}_{\mathcal{T},n}\mathcal{C}_{U,n}}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}}\hat{\mathbf{V}}_n(g)'\hat{\mathbf{V}}_n(g)=\mathcal{C}_{\mathcal{T},n}\mathcal{C}_{U,n}\hat{\Gamma}_n,\end{aligned}$$

we have $\sqrt{n}(\hat{\beta}_n - \beta) = O_p(\sqrt{\mathcal{C}_{\mathcal{T},n}}) \neq O_p(1)$ in general.

- For each $G \in \{1, \ldots, N_{\mathcal{S},n}\}$, let

$$\tilde{\mathbf{r}}_n(G) = (\tilde{r}_{s_{G,n}(1),n}, \ldots, \tilde{r}_{s_{G,n}(\#\mathcal{S}_{G,n}),n})',$$

$$\mathbf{V}_n(G) = (\mathbf{V}_{s_{G,n}(1),n}, \ldots, \mathbf{V}_{s_{G,n}(\#\mathcal{S}_{G,n}),n})',$$

where $s_{G,n}(\cdot)$ is any function such that $\{s_{G,n}(1), \ldots, s_{G,n}(\#\mathcal{S}_{G,n})\} = \mathcal{S}_{G,n}$. Defining

$$\mathcal{C}_{\mathbf{V},n} = \max_{1 \leq G \leq N_{\mathcal{S},n}} \lambda_{\max}(\mathbb{E}[\mathbf{V}_n(G)\mathbf{V}_n(G)'|\mathcal{W}_n]) \leq \mathcal{C}_{\mathcal{S},n}\mathcal{C}_{V,n},$$

the inequality

$$\mathbb{V}\left[\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \tilde{r}_{i,n} V_{i,n} \middle| \mathcal{W}_n\right] \leq \mathcal{C}_{\mathcal{S},n}\mathcal{C}_{V,n}\tilde{\rho}_n$$

can be generalized as follows:

$$\begin{aligned}\mathbb{V}\left[\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \tilde{r}_{i,n} V_{i,n} \middle| \mathcal{W}_n\right] &= \frac{1}{n} \sum_{1 \leq G \leq N_{\mathcal{S},n}} \tilde{\mathbf{r}}_n(G)'\mathbb{E}[\mathbf{V}_n(G)\mathbf{V}_n(G)'|\mathcal{W}_n]\tilde{\mathbf{r}}_n(G) \\ &\leq \frac{\mathcal{C}_{\mathbf{V},n}}{n} \sum_{1 \leq G \leq N_{\mathcal{S},n}} \tilde{\mathbf{r}}_n(G)'\tilde{\mathbf{r}}_n(G) = \mathcal{C}_{\mathbf{V},n}\tilde{\rho}_n.\end{aligned}$$

- Defining

$$\mathcal{C}_{\mathbf{U},n} = \max_{1 \leq g \leq N_{\mathcal{T},n}} \lambda_{\max}(\mathbb{E}[\{\mathbf{U}_n(g)\mathbf{U}_n(g)'\} \otimes \{\mathbf{U}_n(g)\mathbf{U}_n(g)'\}|\mathcal{X}_n, \mathcal{W}_n]) \leq \mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n},$$

the inequality

$$\frac{\lambda_{\min}(\Sigma_n)^2}{n^2} \sum_{1 \leq g \leq N_{\mathcal{T},n}} \mathbb{E}[\eta_n(g)^4|\mathcal{X}_n, \mathcal{W}_n] \leq \frac{\mathcal{C}_{\mathcal{T},n}^3\mathcal{C}_{U,n}}{n^2} \sum_{1 \leq i \leq n} \hat{v}_{i,n}^4$$

can be generalized as follows:

$$\frac{\lambda_{\min}(\Sigma_n)^2}{n^2} \sum_{1\le g\le N_{\mathcal{T},n}} \mathbb{E}[\eta_n(g)^4|\mathcal{X}_n,\mathcal{W}_n] \le \frac{1}{n^2} \sum_{1\le g\le N_{\mathcal{T},n}} \mathbb{E}[(\hat{\mathbf{V}}_n(g)'\mathbf{U}_n(g))^4|\mathcal{X}_n,\mathcal{W}_n]$$

The right hand side can be written as

$$\frac{1}{n^2} \sum_{1\le g\le N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'\mathbb{E}[\{\mathbf{U}_n(g)\mathbf{U}_n(g)'\}\otimes\{\mathbf{U}_n(g)\mathbf{U}_n(g)'\}|\mathcal{X}_n,\mathcal{W}_n]$$

$$(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g)) \le \frac{\mathcal{C}_{\mathbf{U},n}}{n^2} \sum_{1\le g\le N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))$$

$$= \frac{\mathcal{C}_{\mathbf{U},n}}{n^2} \sum_{1\le g\le N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(g)'\hat{\mathbf{V}}_n(g))^2 \le \frac{\mathcal{C}_{\mathcal{T},n}\mathcal{C}_{\mathbf{U},n}}{n^2} \sum_{1\le i\le n} \hat{v}_{i,n}^4,$$

where the last inequality uses

$$(\hat{\mathbf{V}}_n(g)'\hat{\mathbf{V}}_n(g))^2 = (\sum_{t\in\mathcal{T}_{g,n}} \hat{v}_{t,n}^2)^2 \le \mathcal{C}_{\mathcal{T},n} \sum_{t\in\mathcal{T}_{i,n}} \hat{v}_{t,n}^4.$$

- Because $\hat{v}_{i,n} = \tilde{V}_{i,n} + \tilde{Q}_{i,n}$, we have

$$\frac{1}{n^2} \sum_{1\le i\le n} \hat{v}_{i,n}^4 \le \frac{8}{n^2} \sum_{1\le i\le n} \tilde{V}_{i,n}^4 + \frac{8}{n^2} \sum_{1\le i\le n} \tilde{Q}_{i,n}^4,$$

where

$$\frac{1}{n^2} \sum_{1\le i\le n} \tilde{V}_{i,n}^4 = O_p\left(\frac{\mathcal{C}_{\mathcal{S},n}^3}{n}\right)$$

because

$$\begin{aligned}
\mathbb{E}[\tilde{V}_{i,n}^4|\mathcal{W}_n] &= \mathbb{E}\left[\left(\sum_{1\le j\le n} M_{ij,n}V_{j,n}\right)^4\middle|\mathcal{W}_n\right] \\
&= \mathbb{E}\left[\left(\sum_{1\le G\le N_{\mathcal{S},n}}\sum_{s\in\mathcal{S}_{G,n}} M_{is,n}V_{s,n}\right)^4\middle|\mathcal{W}_n\right] \\
&= \sum_{1\le G\le N_{\mathcal{S},n}} \mathbb{E}\left[\left(\sum_{s\in\mathcal{S}_{G,n}} M_{is,n}V_{s,n}\right)^4\middle|\mathcal{W}_n\right] \\
&\quad +3\sum_{1\le G,H\le N_{\mathcal{S},n},G\ne H} \mathbb{E}\left[\left(\sum_{s\in\mathcal{S}_{G,n}} M_{is,n}V_{s,n}\right)^2\left(\sum_{t\in\mathcal{S}_{H,n}} M_{it,n}V_{t,n}\right)^2\middle|\mathcal{W}_n\right] \\
&\le \mathcal{C}_{\mathcal{S},n}^3\mathcal{C}_{V,n}\sum_{1\le G\le N_{\mathcal{S},n}}\sum_{s\in\mathcal{S}_{G,n}} M_{is,n}^4 \\
&+ 3\mathcal{C}_{\mathcal{S},n}^2\mathcal{C}_{V,n}\sum_{1\le G,H\le N_{\mathcal{S},n},G\ne H}\sum_{s\in\mathcal{S}_{G,n},t\in\mathcal{S}_{H,n}} M_{is,n}^2M_{it,n}^2 \\
&\le 3\mathcal{C}_{\mathcal{S},n}^3\mathcal{C}_{V,n}\sum_{1\le j,k\le n} M_{ij,n}^2M_{ik,n}^2 = 3\mathcal{C}_{\mathcal{S},n}^3\mathcal{C}_{V,n}M_{ii,n}^2 \le 3\mathcal{C}_{\mathcal{S},n}^3\mathcal{C}_{V,n},
\end{aligned}$$

and where

$$\frac{1}{n^2}\sum_{1\le i\le n}\tilde{Q}_{i,n}^4 = O_p(\chi_n^2)$$

because

$$\frac{1}{n^2}\sum_{1\le i\le n}\tilde{Q}_{i,n}^4 \le \left(\frac{1}{n}\max_{1\le i\le n}\tilde{Q}_{i,n}^2\right)\left(\frac{1}{n}\sum_{1\le i\le n}\tilde{Q}_{i,n}^2\right) \le \left(\frac{1}{n}\sum_{1\le i\le n}\tilde{Q}_{i,n}^2\right)^2.$$

- To possibly improve the bound

$$\frac{1}{n}\max_{1\le i\le n}\tilde{Q}_{i,n}^2 \le \frac{1}{n}\sum_{1\le i\le n}\tilde{Q}_{i,n}^2 = O_p(\chi_n),$$

suppose that

$$\frac{1}{n}\sum_{1\leq i\leq n}\mathbb{E}[|Q_{i,n}|^{\theta}]=O(1)$$

for some $\theta \geq 2$. Then, by the proof of Lemma SA-7 of the Supplemental Appendix of CJN,

$$\frac{1}{n}\max_{1\leq i\leq n}\tilde{Q}_{i,n}^2=\max\left(n^{-\frac{\theta-2}{\theta}},\mathcal{M}_n\chi_n\right)O_p(1)$$

and

$$\frac{1}{n}\max_{1\leq i\leq n}\tilde{Q}_{i,n}^2=n^{-\frac{\theta-2}{\theta}}\mathcal{C}_{M,n}^{2(\theta-1)/\theta}O_p(1).$$

- The condition

$$\frac{\mathcal{C}_{\mathcal{T},n}^3}{n^2}\sum_{1\leq i\leq n}||\hat{\mathbf{v}}_{i,n}||^4=o_p(1)$$

is satisfied if $\mathcal{C}_{\mathcal{S},n}^3\mathcal{C}_{\mathcal{T},n}^3=o(n)$ and if $\mathcal{C}_{\mathcal{T},n}^{3/2}\chi_n=o(1)$.

- The conditions $\mathcal{C}_{\mathcal{S},n}^3\mathcal{C}_{\mathcal{T},n}^3=o(n)$ and $\mathcal{C}_{\mathcal{S},n}\rho_n+\mathcal{C}_{\mathcal{T},n}^{3/2}\chi_n=o(1)$ are satisfied if one of the following sets of conditions is satisfied:
  - $\mathcal{C}_{\mathcal{S},n}^3\mathcal{C}_{\mathcal{T},n}^3=o(n)$ and $n^{1/3}\rho_n+n^{1/2}\chi_n=O(1)$
  - $\mathcal{C}_{\mathcal{S},n}=\mathcal{C}_{\mathcal{T},n}=o(n^{1/6})$ and $n^{1/6}\rho_n+n^{1/4}\chi_n=O(1)$
  - $\mathcal{C}_{\mathcal{S},n}=O(1),\mathcal{C}_{\mathcal{T},n}=o(n^{1/3}),\rho_n=o(1)$, and $n^{1/2}\chi_n=O(1)$
  - $\mathcal{C}_{\mathcal{S},n}=o(n^{1/3}),\mathcal{C}_{\mathcal{T},n}=O(1),n^{1/3}\rho_n=O(1)$, and $\chi_n=o(1)$

□

Proof of Lemma B.3. It suffices to show that $\tilde{\Sigma}_n=\bar{\Sigma}_n+o_p(1)$ and that $\hat{\Sigma}_n=\tilde{\Sigma}_n+o_p(1)$.

First,

$$\tilde{\Sigma}_n=\frac{1}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}}c_{gg,n}+\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n},g<h}[c_{gh,n}+c_{hg,n}],$$

where $c_{gh,n}$ is

$$c_{gh,n}(\kappa_n) = \sum_{1\leq k,l\leq N_{\mathcal{T},n}} [\hat{\mathbf{V}}_n(k)\otimes\hat{\mathbf{V}}_n(k)]'\kappa_n(k,l)[\mathbf{M}_n(l,g)\otimes\mathbf{M}_n(l,h)][\mathbf{U}_n(g)\otimes\mathbf{U}_n(h)].$$

Defining $\mathcal{C}_{\kappa,n} = ||\kappa_n||_\infty$, we have

$$\begin{aligned}
&\sum_{1\leq g,h\leq N_{\mathcal{T},n}} \mathbb{V}[c_{gh,n}|\mathcal{X}_n,\mathcal{W}_n] \\
\leq\ & \mathcal{C}^2_{\mathcal{T},n}\mathcal{C}_{U,n} \sum_{1\leq k,l,K,L\leq N_{\mathcal{T},n}} [\hat{\mathbf{V}}_n(k)\otimes\hat{\mathbf{V}}_n(k)]' \\
& \kappa_n(k,l)[\mathbf{M}_n(l,L)\otimes\mathbf{M}_n(l,L)]\kappa_n(L,K)[\hat{\mathbf{V}}_n(K)\otimes\hat{\mathbf{V}}_n(K)] \\
=\ & \mathcal{C}^2_{\mathcal{T},n}\mathcal{C}_{U,n}(\hat{\mathbf{V}}_n\circledast_n\hat{\mathbf{V}}_n)'\kappa_n'(\mathbf{M}_n\circledast_n\mathbf{M}_n)\kappa_n(\hat{\mathbf{V}}_n\circledast_n\hat{\mathbf{V}}_n) \\
\leq\ & \mathcal{C}^2_{\kappa,n}\mathcal{C}^2_{\mathcal{T},n}\mathcal{C}_{U,n}||\mathbf{M}_n\circledast_n\mathbf{M}_n||_\infty||\hat{\mathbf{V}}_n\circledast_n\hat{\mathbf{V}}_n||^2 \\
\leq\ & \mathcal{C}^2_{\kappa,n}\mathcal{C}^4_{\mathcal{T},n}\mathcal{C}_{U,n}\sum_{1\leq i\leq n}\hat{v}^4_{i,n} = o_p(n^2),
\end{aligned}$$

where the first inequality uses

$$\sum_{1\leq g,h\leq N_{\mathcal{T},n}} [\mathbf{M}_n(l,g)\mathbf{M}_n(g,L)\otimes\mathbf{M}_n(l,h)\mathbf{M}_n(h,L)] = \mathbf{M}_n(l,L)\otimes\mathbf{M}_n(l,L),$$

the first equality employs the notation

$$\hat{\mathbf{V}}_n\circledast_n\hat{\mathbf{V}}_n = \begin{pmatrix} \hat{\mathbf{V}}_n(1)\otimes\hat{\mathbf{V}}_n(1) \\ \vdots \\ \hat{\mathbf{V}}_n(N_{\mathcal{T},n})\otimes\hat{\mathbf{V}}_n(N_{\mathcal{T},n}) \end{pmatrix},$$

and where the last inequality uses

$$
\begin{aligned}
& ||\mathbf{M}_n \circledast_n \mathbf{M}_n||_\infty \\
= & \max_{1\leq g\leq N_{\mathcal{T},n}} \max_{1\leq k,K\leq \#\mathcal{T}_{g,n}} \sum_{1\leq h\leq N_{\mathcal{T},n}} \left(\sum_{1\leq l\leq \#\mathcal{T}_{h,n}} |M_{t_{g,n}(k),t_{h,n}(l),n}| \right) \\
& \left(\sum_{1\leq L\leq \#\mathcal{T}_{h,n}} |M_{t_{g,n}(K),t_{h,n}(L),n}| \right) \\
\leq & \max_{1\leq g\leq N_{\mathcal{T},n}} \max_{1\leq k,K\leq \#\mathcal{T}_{g,n}} \sum_{1\leq h\leq N_{\mathcal{T},n}} \frac{1}{2} \left[\left(\sum_{1\leq l\leq \#\mathcal{T}_{h,n}} |M_{t_{g,n}(k),t_{h,n}(l),n}| \right)^2 \right. \\
+ & \left. \left(\sum_{1\leq L\leq \#\mathcal{T}_{h,n}} |M_{t_{g,n}(K),t_{h,n}(L),n}| \right)^2 \right] \\
\leq & \max_{1\leq g\leq N_{\mathcal{T},n}} \max_{1\leq k,K\leq \#\mathcal{T}_{g,n}} \sum_{1\leq h\leq N_{\mathcal{T},n}} (\#\mathcal{T}_{h,n}) \\
& \frac{1}{2} \left[\sum_{1\leq l\leq \#\mathcal{T}_{h,n}} M^2_{t_{g,n}(k),t_{h,n}(l),n} + \sum_{1\leq L\leq \#\mathcal{T}_{h,n}} M^2_{t_{g,n}(K),t_{h,n}(L),n} \right] \\
\leq & \mathcal{C}_{\mathcal{T},n} \max_{1\leq g\leq N_{\mathcal{T},n}} \max_{1\leq k\leq \#\mathcal{T}_{g,n}} \sum_{1\leq h\leq N_{\mathcal{T},n}} \sum_{1\leq l\leq \#\mathcal{T}_{h,n}} M^2_{t_{g,n}(k),t_{h,n}(l),n} \\
= & \mathcal{C}_{\mathcal{T},n} \max_{1\leq g\leq N_{\mathcal{T},n}} \max_{1\leq k\leq \#\mathcal{T}_{g,n}} M_{t_{g,n}(k),t_{g,n}(k),n} \leq \mathcal{C}_{\mathcal{T},n}
\end{aligned}
$$

and

$$\begin{aligned}
||\hat{\mathbf{V}}_n \circledast_n \hat{\mathbf{V}}_n||^2 &= \sum_{1\leq g\leq N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g)) \\
&= \sum_{1\leq g\leq N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(g)'\hat{\mathbf{V}}_n(g))^2 = \sum_{1\leq g\leq N_{\mathcal{T},n}} \left(\sum_{t\in\mathcal{T}_{g,n}} \hat{v}_{t,n}^2\right)^2 \\
&\leq \sum_{1\leq g\leq N_{\mathcal{T},n}} (\#\mathcal{T}_{g,n}) \sum_{t\in\mathcal{T}_{g,n}} \hat{v}_{t,n}^4 \leq \mathcal{C}_{\mathcal{T},n} \sum_{1\leq g\leq N_{\mathcal{T},n}} \sum_{t\in\mathcal{T}_{g,n}} \hat{v}_{t,n}^4 \\
&= \mathcal{C}_{\mathcal{T},n} \sum_{1\leq i\leq n} \hat{v}_{i,n}^4.
\end{aligned}$$

As a consequence,

$$\begin{aligned}
\mathbb{V}\left[\frac{1}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}} c_{gg,n}\middle|\mathcal{X}_n,\mathcal{W}_n\right] &= \frac{1}{n^2}\sum_{1\leq g\leq N_{\mathcal{T},n}} \mathbb{V}[c_{gg,n}|\mathcal{X}_n,\mathcal{W}_n] \\
&\leq \frac{1}{n^2}\sum_{1\leq g,h\leq N_{\mathcal{T},n}} \mathbb{V}[c_{gh,n}|\mathcal{X}_n,\mathcal{W}_n] = o_p(1)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{V}\left[\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n},g<h} [c_{gh,n}+c_{hg,n}]\middle|\mathcal{X}_n,\mathcal{W}_n\right] &= \frac{1}{n^2}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}^{g<h} \mathbb{V}[c_{gh,n}+c_{hg,n}|\mathcal{X}_n,\mathcal{W}_n] \\
&\leq \frac{2}{n^2}\sum_{1\leq g,h\leq N_{\mathcal{T},n}} \mathbb{V}[c_{gh,n}|\mathcal{X}_n,\mathcal{W}_n] \\
&= o_p(1),
\end{aligned}$$

implying in particular that $\tilde{\Sigma}_n = \bar{\Sigma}_n + o_p(1)$.

To complete the proof, it therefore suffices to show that $\hat{\Sigma}_n - \tilde{\Sigma}_n = o_p(1)$. Using the Cauchy-Schwarz inequality, simple bounding arguments, and the decompositions

$$\hat{\mathbf{u}}_n(h) - \tilde{\mathbf{U}}_n(h) = \tilde{\mathbf{R}}_n(h) - \hat{\mathbf{V}}_n(h)(\hat{\beta}_n - \beta), \qquad \tilde{\mathbf{R}}_n(h) = \tilde{\mathbf{r}}_n(h) + (\tilde{\mathbf{R}}_n(h) - \tilde{\mathbf{r}}_n(h)),$$

$$\hat{\mathbf{V}}_n(h) = \tilde{\mathbf{V}}_n(h) + \tilde{\mathbf{Q}}_n(h),$$

it can be shown that

$$\begin{aligned}
& \hat{\Sigma}_n - \tilde{\Sigma}_n \\
= \; & \frac{1}{n} \sum_{1 \le g,h \le N_{\mathcal{T},n}} [\hat{\mathbf{V}}_n(g) \otimes \hat{\mathbf{V}}_n(g)]' \kappa_n(g,h) [\hat{\mathbf{u}}_n(h) \otimes \hat{\mathbf{u}}_n(h) - \tilde{\mathbf{U}}_n(h) \otimes \tilde{\mathbf{U}}_n(h)] \\
= \; & \frac{1}{n} \sum_{1 \le g,h \le N_{\mathcal{T},n}} [\hat{\mathbf{V}}_n(g) \otimes \hat{\mathbf{V}}_n(g)]' \kappa_n(g,h) [\{\hat{\mathbf{u}}_n(h) - \tilde{\mathbf{U}}_n(h)\} \otimes \{\hat{\mathbf{u}}_n(h) - \tilde{\mathbf{U}}_n(h)\}] \\
+ \; & \frac{1}{n} \sum_{1 \le g,h \le N_{\mathcal{T},n}} [\hat{\mathbf{V}}_n(g) \otimes \hat{\mathbf{V}}_n(g)]' \kappa_n(g,h) [\tilde{\mathbf{U}}_n(h) \otimes \{\hat{\mathbf{u}}_n(h) - \tilde{\mathbf{U}}_n(h)\} \\
+ \; & \{\hat{\mathbf{u}}_n(h) - \tilde{\mathbf{U}}_n(h)\} \otimes \tilde{\mathbf{U}}_n(h)] = o_p(1)
\end{aligned}$$

if

$$\frac{1}{4n} \sum\nolimits^{\kappa,n} \{(\tilde{V}^2_{t_{g,n}(k),n} + \tilde{V}^2_{t_{g,n}(K),n})(\tilde{r}^2_{t_{h,n}(l),n} + \tilde{r}^2_{t_{h,n}(L),n})\} = O_p(\mathcal{C}_{\mathcal{S},n} \mathcal{C}_{\mathcal{T},n} \rho_n),$$

$$\frac{1}{4n} \sum\nolimits^{\kappa,n} \{(\tilde{V}^2_{t_{g,n}(k),n} + \tilde{V}^2_{t_{g,n}(K),n})[(\tilde{R}_{t_{h,n}(l),n} - \tilde{r}_{t_{h,n}(l),n})^2 + (\tilde{R}_{t_{h,n}(L),n} - \tilde{r}_{t_{h,n}(L),n})^2]\}$$
$$= O_p[\mathcal{C}_{\mathcal{T},n} n(\varrho_n - \rho_n)],$$

$$\frac{1}{4n} \sum\nolimits^{\kappa,n} \{(\tilde{Q}^2_{t_{g,n}(k),n} + \tilde{Q}^2_{t_{g,n}(K),n})(\tilde{R}^2_{t_{h,n}(l),n} + \tilde{R}^2_{t_{h,n}(L),n})\} = O_p(\mathcal{C}_{\mathcal{T},n} n \chi_n \varrho_n),$$

$$\frac{(\hat{\beta}_n - \beta)^2}{4n} \sum\nolimits^{\kappa,n} \{\hat{v}^4_{t_{g,n}(k),n} + \hat{v}^4_{t_{g,n}(K),n} + \hat{v}^4_{t_{h,n}(l),n} + \hat{v}^4_{t_{h,n}(L),n}\} = \left(\frac{\mathcal{C}^2_{\mathcal{T},n}}{n^2} \sum_{1 \le i \le n} \hat{v}^4_{i,n} \right) O_p(1),$$

and if

$$\frac{1}{4n} \sum\nolimits^{\kappa,n} \{(\hat{v}^2_{t_g(k),n} + \hat{v}^2_{t_g(K),n})(\tilde{U}^2_{t_h(l),n} + \tilde{U}^2_{t_h(L),n})\} = O_p(\mathcal{C}^2_{\mathcal{T},n}),$$

where "$\sum^{\kappa,n}\{\cdot\}$" is shorthand for

$$\text{“} \sum_{1 \le g,h \le N_{\mathcal{T},n}} \sum_{1 \le k,K \le \#\mathcal{T}_{g,n}} \sum_{1 \le l,L \le \#\mathcal{T}_{h,n}} |\kappa_{(k-1)(\#\mathcal{T}_{g,n})+K,(l-1)(\#\mathcal{T}_{h,n})+L,n}| \{\cdot\}\text{”}.$$

First,

$$\frac{1}{4n}\sum\nolimits^{\kappa,n}\{(\tilde{V}^2_{t_g(k),n}+\tilde{V}^2_{t_g(K),n})(\tilde{r}^2_{t_h(l),n}+\tilde{r}^2_{t_h(L),n})\}=O_p(\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{\mathcal{T},n}\rho_n)$$

because

$$\begin{aligned}
& \mathbb{E}\left[\frac{1}{4n}\sum\nolimits^{\kappa,n}\{(\tilde{V}^2_{t_g(k),n}+\tilde{V}^2_{t_g(K),n})(\tilde{r}^2_{t_h(l),n}+\tilde{r}^2_{t_h(L),n})\}\middle|\mathcal{W}_n\right] \\
= & \frac{1}{4n}\sum\nolimits^{\kappa,n}\{(\tilde{r}^2_{t_h(l),n}+\tilde{r}^2_{t_h(L),n})\mathbb{E}[(\tilde{V}^2_{t_g(k),n}+\tilde{V}^2_{t_g(K),n})|\mathcal{W}_n]\} \\
\leq & \frac{\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{V,n}}{2n}\sum\nolimits^{\kappa,n}\{\tilde{r}^2_{t_h(l),n}+\tilde{r}^2_{t_h(L),n}\} \\
\leq & \frac{\mathcal{C}_{\kappa,n}\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{\mathcal{T},n}\mathcal{C}_{V,n}}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}}\sum_{t\in\mathcal{T}_{g,n}}\tilde{r}^2_{t,n}=\mathcal{C}_{\kappa,n}\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{\mathcal{T},n}\mathcal{C}_{V,n}\tilde{\rho}_n,
\end{aligned}$$

where the first inequality uses

$$\begin{aligned}
\mathbb{E}[\tilde{V}^2_{i,n}|\mathcal{W}_n] &= \mathbb{E}\left[\left(\sum_{1\leq j\leq n}M_{ij,n}V_{j,n}\right)^2\middle|\mathcal{W}_n\right] \\
&= \mathbb{E}\left[\left(\sum_{1\leq G\leq N_{\mathcal{S},n}}\sum_{s\in\mathcal{S}_{G,n}}M_{is,n}V_{s,n}\right)^2\middle|\mathcal{W}_n\right] \\
&= \sum_{1\leq G\leq N_{\mathcal{S},n}}\mathbb{E}\left[\left(\sum_{s\in\mathcal{S}_{G,n}}M_{is,n}V_{s,n}\right)^2\middle|\mathcal{W}_n\right] \\
&\leq \mathcal{C}_{\mathcal{S},n}\mathcal{C}_{V,n}\sum_{1\leq G\leq N_{\mathcal{S},n}}\sum_{s\in\mathcal{S}_{G,n}}M^2_{is,n}=\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{V,n}M_{ii,n}\leq\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{V,n}.
\end{aligned}$$

Next,

$$\begin{aligned}
& \frac{1}{4n}\sum\nolimits^{\kappa,n}\{(\tilde{V}^2_{t_g(k),n}+\tilde{V}^2_{t_g(K),n})[(\tilde{R}_{t_h(l),n}-\tilde{r}_{t_h(l),n})^2+(\tilde{R}_{t_h(L),n}-\tilde{r}_{t_h(L),n})^2]\} \\
\leq\ & \frac{\mathcal{C}_{\kappa,n}\mathcal{C}_{\mathcal{T},n}}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}\sum_{s\in\mathcal{T}_{g,n}}\sum_{t\in\mathcal{T}_{h,n}}\tilde{V}^2_{s,n}(\tilde{R}_{t,n}-\tilde{r}_{t,n})^2 \\
=\ & n\mathcal{C}_{\kappa,n}\mathcal{C}_{\mathcal{T},n}\left(\frac{1}{n}\sum_{1\leq i\leq n}\tilde{V}^2_{i,n}\right)\left(\frac{1}{n}\sum_{1\leq i\leq n}|\tilde{R}_{i,n}-\tilde{r}_{i,n}|^2\right)=O_p[\mathcal{C}_{\mathcal{T},n}n(\varrho_n-\rho_n)]
\end{aligned}$$

because

$$\frac{1}{n}\sum_{1\leq i\leq n}\tilde{V}^2_{i,n}\leq\frac{1}{n}\sum_{1\leq i\leq n}V^2_{i,n}=\mathcal{C}_{V,n}O_p(1)$$

and

$$\frac{1}{n}\sum_{1\leq i\leq n}|\tilde{R}_{i,n}-\tilde{r}_{i,n}|^2\leq\frac{1}{n}\sum_{1\leq i\leq n}|R_{i,n}-r_{i,n}|^2=O_p(\varrho_n-\rho_n).$$

Similarly, using

$$\tilde{\varrho}_n=\frac{1}{n}\sum_{1\leq i\leq n}\tilde{R}^2_{i,n}\leq\frac{1}{n}\sum_{1\leq i\leq n}R^2_{i,n}=O_p(\varrho_n),$$

we have

$$\begin{aligned}
& \frac{1}{4n}\sum\nolimits^{\kappa,n}\{(\tilde{Q}^2_{t_g(k),n}+\tilde{Q}^2_{t_g(K),n})(\tilde{R}^2_{t_h(l),n}+\tilde{R}^2_{t_h(L),n})\} \\
\leq\ & \frac{\mathcal{C}_{\kappa,n}\mathcal{C}_{\mathcal{T},n}}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}\sum_{s\in\mathcal{T}_{g,n}}\sum_{t\in\mathcal{T}_{h,n}}\tilde{Q}^2_{s,n}\tilde{R}^2_{t,n}=n\mathcal{C}_{\kappa,n}\mathcal{C}_{\mathcal{T},n}\tilde{\chi}_n\tilde{\varrho}_n=O_p(\mathcal{C}_{\mathcal{T},n}n\chi_n\varrho_n).
\end{aligned}$$

Also,

$$\begin{aligned}
& \frac{(\hat{\beta}_n-\beta)^2}{4n}\sum\nolimits^{\kappa,n}\{\hat{v}^4_{t_g(k),n}+\hat{v}^4_{t_g(K),n}+\hat{v}^4_{t_h(l),n}+\hat{v}^4_{t_h(L),n}\} \\
& \qquad =\left(\frac{\mathcal{C}^2_{\mathcal{T},n}}{n^2}\sum_{1\leq i\leq n}\hat{v}^4_{i,n}\right)O_p(1)
\end{aligned}$$

because $n(\hat{\beta}_n - \beta)^2 = O_p(\mathcal{C}_{\mathcal{T},n})$ and

$$\begin{aligned}\frac{\sum^{\kappa,n}\{\hat{v}^4_{t_g(k),n} + \hat{v}^4_{t_g(K),n} + \hat{v}^4_{t_h(l),n} + \hat{v}^4_{t_h(L),n}\}}{4n^2} &\leq \frac{\mathcal{C}_{\kappa,n}\mathcal{C}_{\mathcal{T},n}}{n^2}\sum_{1\leq g\leq N_{\mathcal{T},n}}\sum_{t\in\mathcal{T}_{g,n}}\hat{v}^4_{t,n} \\ &= \mathcal{C}_{\kappa,n}\frac{\mathcal{C}_{\mathcal{T},n}}{n^2}\sum_{1\leq i\leq n}\hat{v}^4_{i,n}.\end{aligned}$$

Finally,

$$\frac{1}{4n}\sum^{\kappa,n}\{(\hat{v}^2_{t_g(k),n} + \hat{v}^2_{t_g(K),n})(\tilde{U}^2_{t_h(l),n} + \tilde{U}^2_{t_h(L),n})\} = O_p(\mathcal{C}^2_{\mathcal{T},n})$$

because

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{4n}\sum^{\kappa,n}\{(\hat{v}^2_{t_g(k),n} + \hat{v}^2_{t_g(K),n})(\tilde{U}^2_{t_h(l),n} + \tilde{U}^2_{t_h(L),n})\}\middle|\mathcal{X}_n,\mathcal{W}_n\right] \\ =\ &\frac{1}{4n}\sum^{\kappa,n}\{(\hat{v}^2_{t_g(k),n} + \hat{v}^2_{t_g(K),n})\mathbb{E}[\tilde{U}^2_{t_h(l),n} + \tilde{U}^2_{t_h(L),n}|\mathcal{X}_n,\mathcal{W}_n]\} \\ \leq\ &\frac{\mathcal{C}_{\mathcal{T},n}\mathcal{C}_{U,n}}{2n}\sum^{\kappa,n}\{(\hat{v}^2_{t_g(k),n} + \hat{v}^2_{t_g(K),n})\} \\ \leq\ &\frac{\mathcal{C}_{\kappa,n}\mathcal{C}^2_{\mathcal{T},n}\mathcal{C}_{U,n}}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}}\sum_{t\in\mathcal{T}_{g,n}}\hat{v}^2_{t,n} = \mathcal{C}_{\kappa,n}\mathcal{C}^2_{\mathcal{T},n}\mathcal{C}_{U,n}\hat{\Gamma}_n = O_p(\mathcal{C}^2_{\mathcal{T},n}),\end{aligned}$$

where the first inequality uses

$$\begin{aligned}\mathbb{E}[\tilde{U}^2_{i,n}|\mathcal{X}_n,\mathcal{W}_n] &= \mathbb{E}\left[\left(\sum_{1\leq j\leq n}M_{ij,n}U_{j,n}\right)^2\middle|\mathcal{X}_n,\mathcal{W}_n\right] \\ &= \mathbb{E}\left[\left(\sum_{1\leq g\leq N_{\mathcal{T},n}}\sum_{s\in\mathcal{T}_{g,n}}M_{is,n}U_{s,n}\right)^2\middle|\mathcal{X}_n,\mathcal{W}_n\right] \\ &= \sum_{1\leq g\leq N_{\mathcal{T},n}}\mathbb{E}\left[\left(\sum_{s\in\mathcal{T}_{g,n}}M_{is,n}U_{s,n}\right)^2\middle|\mathcal{X}_n,\mathcal{W}_n\right] \\ &\leq \mathcal{C}_{\mathcal{T},n}\mathcal{C}_{U,n}\sum_{1\leq g\leq N_{\mathcal{T},n}}\sum_{s\in\mathcal{T}_{g,n}}M^2_{is,n} = \mathcal{C}_{\mathcal{T},n}\mathcal{C}_{U,n}M_{ii,n} \leq \mathcal{C}_{\mathcal{T},n}\mathcal{C}_{U,n}.\end{aligned}$$

Remark B.3.

- The following assumptions are used in the proof:
  - $\max_{1\leq i\leq n}\mathbb{E}[U_{i,n}^4|\mathcal{X}_n,\mathcal{W}_n]=O_p(1)$
  - $\max_{1\leq i\leq n}\mathbb{E}[\|\mathbf{V}_{i,n}\|^4|\mathcal{W}_n]=O_p(1)$
  - $\chi_n=O(1)$
  - $\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{\mathcal{T},n}^3\rho_n=o(1)$
  - $\mathcal{C}_{\mathcal{T},n}^3 n(\varrho_n-\rho_n)=o(1)$
  - $\mathcal{C}_{\mathcal{T},n}^3 n\chi_n\varrho_n=o(1)$
  - $\mathcal{C}_{\mathcal{T},n}^4 n^{-2}\sum_{1\leq i\leq n}||\hat{\mathbf{v}}_{i,n}||^4=o_p(1)$
  - $||\kappa_n||_\infty=O_p(1)$
- The proof makes repeated use of the following fact: By the triangle and Cauchy-Schwarz inequalities, we have (in generic, but obvious, notation)

$$
\begin{aligned}
&\left|\frac{1}{4n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}[\mathbf{x}_n(g)\otimes\mathbf{X}_n(g)]'\kappa_n(g,h)[\mathbf{y}_n(h)\otimes\mathbf{Y}_n(h)]\right|\\
\leq\;&\frac{1}{4n}\sum^{\kappa,n}\{|x_{t_{g,n}(k),n}X_{t_{g,n}(K),n}y_{t_{h,n}(l),n}Y_{t_{h,n}(L),n}|\}\\
\leq\;&\sqrt{\frac{1}{4n}\sum^{\kappa,n}\{x^2_{t_{g,n}(k),n}y^2_{t_{h,n}(l),n}\}}\sqrt{\frac{1}{4n}\sum^{\kappa,n}\{X^2_{t_{g,n}(K),n}Y^2_{t_{h,n}(L),n}\}}\\
\leq\;&\sqrt{\frac{1}{4n}\sum^{\kappa,n}\{(x^2_{t_{g,n}(k),n}+x^2_{t_{g,n}(K),n})(y^2_{t_{h,n}(l),n}+y^2_{t_{h,n}(L),n})\}}\\
&\sqrt{\frac{1}{4n}\sum^{\kappa,n}\{(X^2_{t_{g,n}(k),n}+X^2_{t_{g,n}(K),n})(Y^2_{t_{h,n}(l),n}+Y^2_{t_{h,n}(L),n})\}}
\end{aligned}
$$

and

$$\begin{aligned}
&\left|\frac{1}{4n}\sum_{1\le g,h\le N_{\mathcal{T},n}}[\mathbf{x}_n(g)\otimes\mathbf{X}_n(g)]'\kappa_n(g,h)[\mathbf{y}_n(h)\otimes\mathbf{Y}_n(h)]\right| \\
\le\ &\sqrt{\frac{1}{4n}\sum{}^{\kappa,n}\{(x^2_{t_{g,n}(k),n}+x^2_{t_{g,n}(K),n})(Y^2_{t_{h,n}(l),n}+Y^2_{t_{h,n}(L),n})\}} \\
&\sqrt{\frac{1}{4n}\sum{}^{\kappa,n}\{(X^2_{t_{g,n}(k),n}+X^2_{t_{g,n}(K),n})(y^2_{t_{h,n}(l),n}+y^2_{t_{h,n}(L),n})\}}.
\end{aligned}$$

- The proof uses

$$\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{\mathcal{T},n}\rho_n+\mathcal{C}_{\mathcal{T},n}n\chi_n\varrho_n+\mathcal{C}_{\mathcal{T},n}n(\varrho_n-\rho_n)=o(1)$$

to show that

$$\frac{1}{n}\sum_{1\le g,h\le N_{\mathcal{T},n}}[\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g)]'\kappa_n(g,h)[\tilde{\mathbf{R}}_n(h)\otimes\tilde{\mathbf{R}}_n(h)]=o_p(1).$$

The stronger condition

$$\mathcal{C}^2_{\mathcal{T},n}[\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{\mathcal{T},n}\rho_n+\mathcal{C}_{\mathcal{T},n}n\chi_n\varrho_n+\mathcal{C}_{\mathcal{T},n}n(\varrho_n-\rho_n)]=o(1)$$

is used (only) to show that

$$\begin{aligned}
&\frac{\sum_{1\le g,h\le N_{\mathcal{T},n}}[\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g)]'\kappa_n(g,h)[\tilde{\mathbf{U}}_n(h)\otimes\tilde{\mathbf{R}}_n(h)+\tilde{\mathbf{R}}_n(h)\otimes\tilde{\mathbf{U}}_n(h)]}{n} \\
&\qquad=o_p(1).
\end{aligned}$$

The extra $\mathcal{C}^2_{\mathcal{T},n}$ is used to control the second term in the bound

$$\begin{aligned}
&\left|\frac{\sum_{1\leq g,h\leq N_{\mathcal{T},n}}[\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g)]'\kappa_n(g,h)[\tilde{\mathbf{U}}_n(h)\otimes\tilde{\mathbf{R}}_n(h)+\tilde{\mathbf{R}}_n(h)\otimes\tilde{\mathbf{U}}_n(h)]}{8n}\right| \\
\leq\ & \sqrt{\frac{1}{4n}\sum\nolimits^{\kappa,n}\{(\hat{v}^2_{t_{g,n}(k),n}+\hat{v}^2_{t_{g,n}(K),n})(\tilde{R}^2_{t_{h,n}(l),n}+\tilde{R}^2_{t_{h,n}(L),n})\}} \\
& \sqrt{\frac{1}{4n}\sum\nolimits^{\kappa,n}\{(\hat{v}^2_{t_{g,n}(k),n}+\hat{v}^2_{t_{g,n}(K),n})(\tilde{U}^2_{t_{h,n}(l),n}+\tilde{U}^2_{t_{h,n}(L),n})\}},
\end{aligned}$$

where the inequality uses the previous remark.

- It is unclear whether a better bound on the magnitude of

$$\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}[\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g)]'\kappa_n(g,h)[\tilde{\mathbf{U}}_n(h)\otimes\tilde{\mathbf{R}}_n(h)+\tilde{\mathbf{R}}_n(h)\otimes\tilde{\mathbf{U}}_n(h)]$$

can be obtained by exploiting the fact that $\mathbb{E}[U_{i,n}|\mathcal{X}_n,\mathcal{W}_n]=0$. Using the representation

$$\tilde{U}_{t_{h,n}(L),n}=\sum_{1\leq m\leq N_{\mathcal{T},n}}\sum_{1\leq M\leq\#\mathcal{T}_{m,n}}M_{t_{h,n}(L)t_{m,n}(M),n}U_{t_{m,n}(M),n},$$

the term

$$\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}[\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g)]'\kappa_n\left(g,h\right)[\tilde{\mathbf{R}}_n(h)\otimes\tilde{\mathbf{U}}_n(h)]$$

can be shown to have (conditional mean zero and) conditional variance bounded

by

$$\begin{aligned}
&\frac{\mathcal{C}_{\mathcal{T},n}\mathcal{C}_{U,n}}{n^2}\sum_{1\leq m\leq N_{\mathcal{T},n}}\sum_{1\leq M\leq \#\mathcal{T}_{m,n}}\\
&\left(\sum\nolimits^{\kappa,n}\{|\hat{v}_{t_{g,n}(k),n}\hat{v}_{t_{g,n}(K),n}\tilde{R}_{t_{h,n}(l),n}\bar{M}_{t_{h,n}(L)t_{m,n}(M),n}|\}\right)^2\\
\leq\ & \mathcal{C}_{\mathcal{T},n}\mathcal{C}_{U,n}\left(\frac{1}{n}\sum\nolimits^{\kappa,n}\hat{v}^2_{t_{i,n}(k),n}\tilde{R}^2_{t_{j,n}(l),n}\right)\\
&\left(\frac{1}{n}\sum_{1\leq m\leq N_{\mathcal{T},n}}\sum_{1\leq M\leq \#\mathcal{T}_{m,n}}\sum\nolimits^{\kappa,n}\hat{v}^2_{t_{i,n}(K),n}M^2_{t_{j,n}(L)t_{m,n}(M),n}\right)\\
=\ & \mathcal{C}_{\mathcal{T},n}\mathcal{C}_{U,n}\left(\frac{1}{n}\sum\nolimits^{\kappa,n}\hat{v}^2_{t_{g,n}(k),n}\tilde{R}^2_{t_{h,n}(l),n}\right)\left(\frac{1}{n}\sum\nolimits^{\kappa,n}\hat{v}^2_{t_{g,n}(K),n}M_{t_{h,n}(L)t_{h,n}(L),n}\right),
\end{aligned}$$

where

$$\frac{1}{n}\sum\nolimits^{\kappa,n}\hat{v}^2_{t_g(k),n}\tilde{R}^2_{t_h(l),n}=O_p[\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{\mathcal{T},n}\rho_n+\mathcal{C}_{\mathcal{T},n}n(\varrho_n-\rho_n)+\mathcal{C}_{\mathcal{T},n}n\chi_n\varrho_n]$$

and

$$\begin{aligned}
\frac{1}{n}\sum\nolimits^{\kappa,n}\hat{v}^2_{t_{g,n}(K),n}M_{t_{h,n}(L)t_{h,n}(L),n}&\leq\frac{\mathcal{C}_{\kappa,n}\mathcal{C}_{\mathcal{T},n}}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}}\sum_{t\in\mathcal{T}_{g,n}}\hat{v}^2_{t,n}\\
&=\mathcal{C}_{\kappa,n}\mathcal{C}_{\mathcal{T},n}\hat{\Gamma}_n=O_p(\mathcal{C}_{\mathcal{T},n}).
\end{aligned}$$

As a consequence, we once again find that

$$\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}[\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g)]'\kappa_n(g,h)[\tilde{\mathbf{U}}_n(h)\otimes\tilde{\mathbf{R}}_n(h)+\tilde{\mathbf{R}}_n(h)\otimes\tilde{\mathbf{U}}_n(h)]=o_p(1)$$

if

$$\mathcal{C}^2_{\mathcal{T},n}[\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{\mathcal{T},n}\rho_n+\mathcal{C}_{\mathcal{T},n}n\chi_n\varrho_n+\mathcal{C}_{\mathcal{T},n}n(\varrho_n-\rho_n)]=o(1).$$

- The condition

$$\frac{\mathcal{C}^4_{\mathcal{T},n}}{n^2}\sum_{1\leq i\leq n}||\hat{\mathbf{v}}_{i,n}||^4=o_p(1)$$

is satisfied if $\mathcal{C}^3_{\mathcal{S},n}\mathcal{C}^4_{\mathcal{T},n}=o(n)$ and if $\mathcal{C}^2_{\mathcal{T},n}\chi_n=o(1)$

- The conditions $\mathcal{C}_{\mathcal{S},n}^3\mathcal{C}_{\mathcal{T},n}^4 = o(n)$ and $\mathcal{C}_{\mathcal{S},n}\mathcal{C}_{\mathcal{T},n}^3\rho_n + \mathcal{C}_{\mathcal{T},n}^3 n\chi_n\varrho_n + \mathcal{C}_{\mathcal{T},n}^3 n(\varrho_n - \rho_n) + \mathcal{C}_{\mathcal{T},n}^2\chi_n = o(1)$ are satisfied if one of the following sets of conditions is satisfied:
  - $\mathcal{C}_{\mathcal{S},n}^3\mathcal{C}_{\mathcal{T},n}^4 = o(n)$ and $n^{5/6}\rho_n + n^{7/4}\chi_n\varrho_n + n^{7/4}(\varrho_n - \rho_n) + n^{1/2}\chi_n = O(1)$
  - $\mathcal{C}_{\mathcal{S},n} = \mathcal{C}_{\mathcal{T},n} = o(n^{1/7})$ and $n^{4/7}\rho_n + n^{10/7}\chi_n\varrho_n + n^{10/7}(\varrho_n - \rho_n) + n^{2/7}\chi_n = O(1)$
  - $\mathcal{C}_{\mathcal{S},n} = O(1), \mathcal{C}_{\mathcal{T},n} = o(n^{1/4})$ and $n^{3/4}\rho_n + n^{7/4}\chi_n\varrho_n + n^{7/4}(\varrho_n - \rho_n) + n^{1/2}\chi_n = O(1)$
  - $\mathcal{C}_{\mathcal{S},n} = o(n^{1/3}), \mathcal{C}_{\mathcal{T},n} = O(1), n^{1/3}\rho_n = O(1)$, and $n\chi_n\varrho_n = o(1) + n(\varrho_n - \rho_n) + \chi_n = o(1)$

□

Proof of Lemma B.4. Defining

$$\mathbf{D}_n = \mathbf{D}_n(\kappa_n) = \kappa_n(\mathbf{M}_n \circledast_n \mathbf{M}_n) - \mathbf{I}_{N_{\kappa,n}}$$

and employing the notation

$$\mathbf{U}_n \circledast_n \mathbf{U}_n = \begin{pmatrix} \mathbf{U}_n(1) \otimes \mathbf{U}_n(1) \\ \vdots \\ \mathbf{U}_n(N_{\mathcal{T},n}) \otimes \mathbf{U}_n(N_{\mathcal{T},n}) \end{pmatrix},$$

we have

$$\bar{\mathbf{\Sigma}}_n - \mathbf{\Sigma}_n = \frac{1}{n}(\hat{\mathbf{V}}_n \circledast_n \hat{\mathbf{V}}_n)'\mathbf{D}_n\mathbb{E}[\mathbf{U}_n \circledast_n \mathbf{U}_n | \mathcal{X}_n, \mathcal{W}_n],$$

so if $\mathcal{C}_{\mathcal{T},n}||\mathbf{D}_n||_\infty = o_p(1)$, then

$$\begin{aligned} |\bar{\mathbf{\Sigma}}_n - \mathbf{\Sigma}_n| &\leq \left\| \frac{1}{n}(\hat{\mathbf{V}}_n \circledast_n \hat{\mathbf{V}}_n) \right\|_1 ||\mathbf{D}_n||_\infty ||\mathbb{E}[\mathbf{U}_n \circledast_n \mathbf{U}_n | \mathcal{X}_n, \mathcal{W}_n]||_\infty \\ &\leq \mathcal{C}_{\mathcal{T},n}\mathcal{C}_{U,n}\hat{\Gamma}_n||\mathbf{D}_n||_\infty = o_p(1), \end{aligned}$$

where $\|\cdot\|_1$ denotes the largest column sum of its argument and where the second

inequality uses

$$\left\|\hat{\mathbf{V}}_n(g) \otimes \hat{\mathbf{V}}_n(g)\right\|_1 = \left\|\hat{\mathbf{V}}_n(g)\right\|_1^2 \leq \mathcal{C}_{\mathcal{T},n} \left\|\hat{\mathbf{V}}_n(g)\right\|^2 = \mathcal{C}_{\mathcal{T},n} \hat{\mathbf{V}}_n(g)' \hat{\mathbf{V}}_n(g),$$

and

$$||\mathbb{E}[\mathbf{U}_n \circledast_n \mathbf{U}_n | \mathcal{X}_n, \mathcal{W}_n]||_\infty \leq \mathcal{C}_{U,n}.$$

Remark B.4.

- The following assumptions are used in the proof:
  - $\max_{1 \leq i \leq n} \mathbb{E}[U_{i,n}^4 | \mathcal{X}_n, \mathcal{W}_n] = O_p(1)$
  - $\max_{1 \leq i \leq n} \mathbb{E}[\|\mathbf{V}_{i,n}\|^4 | \mathcal{W}_n] = O_p(1)$
  - $\chi_n = O(1)$

□

Proof of Lemma B.5. For any $N_{\kappa,n} \times N_{\kappa,n}$ matrix $\mathbf{D}_n$ partitioned conformably with $\mathbf{M}_n \circledast_n \mathbf{M}_n$ as

$$\mathbf{D}_n = \begin{pmatrix} \mathbf{D}_n(1,1) & \mathbf{D}_n(1,2) & \cdots & \mathbf{D}_n(1, N_{\mathcal{T},n}) \\ \mathbf{D}_n(2,1) & \mathbf{D}_n(2,2) & \cdots & \mathbf{D}_n(2, N_{\mathcal{T},n}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{D}_n(N_{\mathcal{T},n}, 1) & \mathbf{D}_n(N_{\mathcal{T},n}, 2) & \cdots & \mathbf{D}_n(N_{\mathcal{T},n}, N_{\mathcal{T},n}) \end{pmatrix},$$

let

$$\begin{aligned} \operatorname{diag}_n(\mathbf{D}_n) &= \mathbf{D}_n(1,1) \oplus \mathbf{D}_n(2,2) \oplus \ldots \oplus \mathbf{D}_n(N_{\mathcal{T},n}, N_{\mathcal{T},n}) \\ &= \begin{pmatrix} \mathbf{D}_n(1,1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_n(2,2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{D}_n(N_{\mathcal{T},n}, N_{\mathcal{T},n}) \end{pmatrix} \end{aligned}$$

and

$$\begin{aligned}\mathrm{diag}_n^{\perp}(\mathbf{D}_n) &= \mathbf{D}_n - \mathrm{diag}_n(\mathbf{D}_n) \\ &= \begin{pmatrix} \mathbf{0} & \mathbf{D}_n(1,2) & \cdots & \mathbf{D}_n(1,N_{\mathcal{T},n}) \\ \mathbf{D}_n(2,1) & \mathbf{0} & \cdots & \mathbf{D}_n(2,N_{\mathcal{T},n}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{D}_n(N_{\mathcal{T},n},1) & \mathbf{D}_n(N_{\mathcal{T},n},2) & \cdots & \mathbf{0} \end{pmatrix}.\end{aligned}$$

For any $g \in \{1,\ldots,N_{\mathcal{T},n}\}$ and any $k,K \in \{1,\ldots,\#\mathcal{T}_{g,n}\}$, we have

$$\begin{aligned}&\sum_{1\le h\le N_{\mathcal{T},n}} \sum_{1\le l,L\le \#\mathcal{T}_{h,n}} \mathbb{I}\left(g\neq h\right) |M_{t_{g,n}(k),t_{h,n}(l),n} M_{t_{g,n}(K),t_{h,n}(L),n}| \\ \le\ & \frac{1}{2} \sum_{1\le h\le N_{\mathcal{T},n}} \sum_{1\le l,L\le \#\mathcal{T}_{h,n}} \mathbb{I}\left(g\neq h\right) \left[M^2_{t_{g,n}(k),t_{h,n}(l),n} + M^2_{t_{g,n}(K),t_{h,n}(L),n}\right] \\ \le\ & \frac{\mathcal{C}_{\mathcal{T},n}}{2} \sum_{1\le h\le N_{\mathcal{T},n}} \sum_{1\le l\le \#\mathcal{T}_{h,n}} \mathbb{I}\left(g\neq h\right) \left[M^2_{t_{g,n}(k),t_{h,n}(l),n} + M^2_{t_{g,n}(K),t_{h,n}(l),n}\right] \\ \le\ & \frac{\mathcal{C}_{\mathcal{T},n}}{2} \left[M_{t_{g,n}(k),t_{g,n}(k),n}(1 - M_{t_{g,n}(k),t_{g,n}(k),n}) \right. \\ +\ & \left. M_{t_{g,n}(K),t_{g,n}(K),n}(1 - M_{t_{g,n}(K),t_{g,n}(K),n})\right] \le \mathcal{C}_{\mathcal{T},n}\mathcal{M}_n.\end{aligned}$$

As a consequence,

$$\begin{aligned}&||\mathrm{diag}_n^{\perp}(\mathbf{M}_n \circledast_n \mathbf{M}_n)||_\infty = \max_{1\le g\le N_{\mathcal{T},n}} \max_{1\le k,K\le \#\mathcal{T}_{g,n}} \\ &\left\{ \sum_{1\le h\le N_{\mathcal{T},n}} \sum_{1\le l,L\le \#\mathcal{T}_{h,n}} \mathbb{I}(g\neq h) |M_{t_{g,n}(k),t_{h,n}(l),n} M_{t_{g,n}(K),t_{h,n}(L),n}| \right\} \\ \le\ & \mathcal{C}_{\mathcal{T},n}\mathcal{M}_n,\end{aligned}$$

a fact that we shall use repeatedly in what follows.

(a) Suppose $\kappa_n = \mathbf{I}_{N_{\kappa,n}}$. Then

$$\begin{aligned}
||\mathbf{D}_n(\kappa_n)||_\infty &\leq ||\text{diag}_n(\mathbf{M}_n \circledast_n \mathbf{M}_n - \mathbf{I}_{N_{\kappa,n}})||_\infty + ||\text{diag}_n^\perp(\mathbf{M}_n \circledast_n \mathbf{M}_n)||_\infty \\
&\leq \max_{1\leq g\leq N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g) \otimes \mathbf{M}_n(g,g) - \mathbf{I}_{\#\mathcal{T}_{g,n}} \otimes \mathbf{I}_{\#\mathcal{T}_{g,n}}||_\infty + \mathcal{C}_{\mathcal{T},n}\mathcal{M}_n \\
&\leq 2(1+\mathcal{C}_{\mathcal{T},n})\mathcal{M}_n,
\end{aligned}$$

where the last inequality uses

$$\begin{aligned}
&||\mathbf{M}_n(g,g) \otimes \mathbf{M}_n(g,g) - \mathbf{I}_{\#\mathcal{T}_{g,n}} \otimes \mathbf{I}_{\#\mathcal{T}_{g,n}}||_\infty \\
\leq\ & \max_{1\leq h\leq N_{\mathcal{T},n}} \max_{1\leq k,K\leq \#\mathcal{T}_{h,n}} \{1 - M_{t_{h,n}(k),t_{h,n}(k),n} M_{t_{h,n}(K),t_{h,n}(K),n}\} + \max_{1\leq h\leq N_{\mathcal{T},n}} \max_{1\leq k,K\leq \#\mathcal{T}_{h,n}} \\
& \left\{ \sum_{1\leq l,L\leq \#\mathcal{T}_{h,n}} \{1 - \mathbb{I}(k=l, K=L)\} |M_{t_{h,n}(k),t_{h,n}(l),n} M_{t_{h,n}(K),t_{h,n}(L),n}| \right\}
\end{aligned}$$

and the fact that

$$\begin{aligned}
1 - M_{t_{h,n}(k),t_{h,n}(k),n} M_{t_{h,n}(K),t_{h,n}(K),n} &\leq 1 - \left(\min_{1\leq i\leq n} M_{ii,n}\right)^2 \\
&= \left(1 + \min_{1\leq i\leq n} M_{ii,n}\right)\mathcal{M}_n \leq 2\mathcal{M}_n
\end{aligned}$$

and

$$\begin{aligned}
& \sum_{1\leq l,L\leq \#\mathcal{T}_{h,n}} \{1 - \mathbb{I}(k=l, K=L)\} |M_{t_{h,n}(k),t_{h,n}(l),n} M_{t_{h,n}(K),t_{h,n}(L),n}| \\
\leq\ & \frac{1}{2} \sum_{1\leq l,L\leq \#\mathcal{T}_{h,n}} \left[\mathbb{I}(k\neq l) M^2_{t_{h,n}(k),t_{h,n}(l),n} + \mathbb{I}(K\neq L) M^2_{t_{h,n}(K),t_{h,n}(L),n}\right] \\
\leq\ & \frac{\mathcal{C}_{\mathcal{T},n}}{2} \sum_{1\leq l\leq \#\mathcal{T}_{h,n}} \left[\mathbb{I}(k\neq l) M^2_{t_{h,n}(k),t_{h,n}(l),n} + \mathbb{I}(K\neq l) M^2_{t_{h,n}(K),t_{h,n}(l),n}\right] \\
\leq\ & \frac{\mathcal{C}_{\mathcal{T},n}}{2} \left[M_{t_{h,n}(k),t_{h,n}(k),n}(1 - M_{t_{h,n}(k),t_{h,n}(k),n}) \right. \\
+\ & \left. M_{t_{h,n}(K),t_{h,n}(K),n}(1 - M_{t_{h,n}(K),t_{h,n}(K),n})\right] \leq \mathcal{C}_{\mathcal{T},n}\mathcal{M}_n.
\end{aligned}$$

This establishes part (a).

(b) Suppose $||\mathbf{M}_n(g,g)^{-1/2}||_\infty^{-2} > 0$ for every $g = 1, \ldots, N_{\mathcal{T},n}$ and that $\kappa_n = \kappa_n^{\mathtt{BR}}$. Then

$$\begin{aligned} ||\kappa_n||_\infty &= \max_{1\leq g\leq N_{\mathcal{T},n}} \left\{||\mathbf{M}_n(g,g)^{-1/2} \otimes \mathbf{M}_n(g,g)^{-1/2}||_\infty\right\} \\ &= \max_{1\leq g\leq N_{\mathcal{T},n}} \left\{||\mathbf{M}_n(g,g)^{-1/2}||_\infty^2\right\} \\ &= \frac{1}{\min_{1\leq g\leq N_{\mathcal{T},n}} \left\{||\mathbf{M}_n(g,g)^{-1/2}||_\infty^{-2}\right\}}. \end{aligned}$$

Also, for any $g \in \{1, \ldots, N_{\mathcal{T},n}\}$ and any $k, K \in \{1, \ldots, \#\mathcal{T}_{g,n}\}$, we have

$$\left(1 - M_{t_{g,n}(k),t_{g,n}(k),n} M_{t_{g,n}(K),t_{g,n}(K),n}\right)^2 \leq 4\mathcal{M}_n^2$$

and

$$\begin{aligned} &\sum_{1\leq l,L\leq \#\mathcal{T}_{g,n}} \{1 - \mathbb{I}\,(k = l, K = L)\} M^2_{t_{g,n}(k),t_{g,n}(l),n} M^2_{t_{g,n}(K),t_{g,n}(L),n} \\ \leq\ & M_{t_{g,n}(k),t_{g,n}(k),n}(1 - M_{t_{g,n}(k),t_{g,n}(k),n}) M_{t_{g,n}(K),t_{g,n}(K),n}(1 - M_{t_{g,n}(K),t_{g,n}(K),n}) \\ \leq\ & \mathcal{M}_n^2, \end{aligned}$$

so

$$\begin{aligned} &||\mathbf{M}_n(g,g) \otimes \mathbf{M}_n(g,g) - \mathbf{I}_{\#\mathcal{T}_{g,n}} \otimes \mathbf{I}_{\#\mathcal{T}_{g,n}}||_F^2 \\ =\ & \sum_{1\leq k,K\leq \#\mathcal{T}_{g,n}} \left(1 - M_{t_{g,n}(k),t_{g,n}(k),n} M_{t_{g,n}(K),t_{g,n}(K),n}\right)^2 \\ &+ \sum_{1\leq k,K,l,L\leq \#\mathcal{T}_{g,n}} \{1 - \mathbb{I}\,(k = l, K = L)\} M^2_{t_{g,n}(k),t_{g,n}(l),n} M^2_{t_{g,n}(K),t_{g,n}(L),n} \\ \leq\ & 5\mathcal{C}^2_{\mathcal{T},n}\mathcal{M}_n^2, \end{aligned}$$

where $||\cdot||_F$ denotes the Frobenius norm. As a consequence, letting $||\cdot||_2$ denote the

spectral norm and using Lemma 2.2 of Schmitt (1992),

$$\begin{aligned}
& ||\mathbf{M}_n(g,g)^{1/2} \otimes \mathbf{M}_n(g,g)^{1/2} - \mathbf{I}_{\#\mathcal{T}_{g,n}} \otimes \mathbf{I}_{\#\mathcal{T}_{g,n}}||_\infty \\
\leq\ & \mathcal{C}_{\mathcal{T},n}||\mathbf{M}_n(g,g)^{1/2} \otimes \mathbf{M}_n(g,g)^{1/2} - \mathbf{I}_{\#\mathcal{T}_{g,n}} \otimes \mathbf{I}_{\#\mathcal{T}_{g,n}}||_2 \\
\leq\ & \mathcal{C}_{\mathcal{T},n}||\mathbf{M}_n(g,g) \otimes \mathbf{M}_n(g,g) - \mathbf{I}_{\#\mathcal{T}_{g,n}} \otimes \mathbf{I}_{\#\mathcal{T}_{g,n}}||_2 \\
\leq\ & \mathcal{C}_{\mathcal{T},n}||\mathbf{M}_n(g,g) \otimes \mathbf{M}_n(g,g) - \mathbf{I}_{\#\mathcal{T}_{g,n}} \otimes \mathbf{I}_{\#\mathcal{T}_{g,n}}||_F \\
\leq\ & \sqrt{5}\mathcal{C}_{\mathcal{T},n}^2 \mathcal{M}_n,
\end{aligned}$$

and therefore

$$\begin{aligned}
& ||\mathbf{D}_n(\kappa_n)||_\infty \\
\leq\ & ||\mathrm{diag}_n[\kappa_n(\mathbf{M}_n \circledast_n \mathbf{M}_n) - \mathbf{I}_{N_{\kappa,n}}]||_\infty + ||\kappa_n \mathrm{diag}_n^\perp(\mathbf{M}_n \circledast_n \mathbf{M}_n)||_\infty \\
\leq\ & \max_{1\leq g\leq N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{1/2} \otimes \mathbf{M}_n(g,g)^{1/2} - \mathbf{I}_{\#\mathcal{T}_{g,n}} \otimes \mathbf{I}_{\#\mathcal{T}_{g,n}}||_\infty \\
+\ & ||\kappa_n||_\infty ||\mathrm{diag}_n^\perp(\mathbf{M}_n \circledast_n \mathbf{M}_n)||_\infty \leq \left(\sqrt{5}\mathcal{C}_{\mathcal{T},n}^2 + ||\kappa_n||_\infty \mathcal{C}_{\mathcal{T},n}\right) \mathcal{M}_n.
\end{aligned}$$

Part (b) now follows from the fact that

$$\mathbb{P}\left[\min_{1\leq g\leq N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{-1/2}||_\infty^{-2} > 0\right] \to 1$$

and

$$\frac{1}{\min_{1\leq g\leq N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{-1/2}||_\infty^{-2}} = O_p(1).$$

(c) Suppose $||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} > 0$ for every $g = 1, \ldots, N_{\mathcal{T},n}$ and that $\kappa_n = \kappa_n^{\mathtt{JN}}$. Then

$$\begin{aligned}
||\kappa_n||_\infty &= \max_{1\leq g\leq N_{\mathcal{T},n}} \left\{||\mathbf{M}_n(g,g)^{-1} \otimes \mathbf{M}_n(g,g)^{-1}||_\infty\right\} = \max_{1\leq g\leq N_{\mathcal{T},n}} \left\{||\mathbf{M}_n(g,g)^{-1}||_\infty^2\right\} \\
&= \frac{1}{\min_{1\leq g\leq N_{\mathcal{T},n}} \left\{||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2}\right\}}.
\end{aligned}$$

Also,

$$\mathbf{D}_n(\kappa_n) = \kappa_n \mathrm{diag}_n^\perp(\mathbf{M}_n \circledast_n \mathbf{M}_n)$$

satisfies

$$\mathbb{E}[vec(\tilde{\mathbf{\Sigma}}_n - \mathbf{\Sigma}_n)|\mathcal{X}_n, \mathcal{W}_n] = \frac{1}{n}(\hat{\mathbf{V}}_n \otimes_n \hat{\mathbf{V}}_n)'\mathbf{D}_n(\kappa_n)\mathbb{E}[\mathbf{U}_n \circledast_n \mathbf{U}_n|\mathcal{X}_n, \mathcal{W}_n] \geq 0$$

and

$$||\mathbf{D}_n(\kappa_n)||_\infty \leq ||\kappa_n||_\infty ||\text{diag}_n^\perp(\mathbf{M}_n \circledast_n \mathbf{M}_n)||_\infty \leq ||\kappa_n||_\infty \mathcal{C}_{\mathcal{T},n}\mathcal{M}_n.$$

Finally, suppose

$$\mathcal{C}_{\mathcal{T},n}\mathcal{M}_n < 1.$$

Then each $\mathbf{M}_n(g,g)$ is diagonally dominant: If $k \in \{1, \ldots, \#\mathcal{T}_{g,n}\}$, then

$$\begin{aligned}
\sum_{1\leq K\leq \#\mathcal{T}_{g,n}, K\neq k} |M_{t_{g,n}(k),t_{g,n}(K),n}| &\leq \sqrt{(\#\mathcal{T}_{g,n}-1)\sum_{1\leq K\leq \#\mathcal{T}_{g,n}, K\neq k} M^2_{t_{g,n}(k),t_{g,n}(K),n}} \\
&\leq \sqrt{(\mathcal{C}_{\mathcal{T},n}-1)M_{t_{g,n}(k),t_{g,n}(k),n}(1-M_{t_{g,n}(k),t_{g,n}(k),n})},
\end{aligned}$$

so

$$\begin{aligned}
& M_{t_{g,n}(k),t_{g,n}(k),n} - \sum_{1\leq K\leq \#\mathcal{T}_{g,n}, K\neq k} |M_{t_{g,n}(k),t_{g,n}(K),n}| \\
\geq & \min_{1\leq i\leq n}\left\{M_{ii,n} - \sqrt{(\mathcal{C}_{\mathcal{T},n}-1)M_{ii,n}(1-M_{ii,n})}\right\} \\
= & 1 - \mathcal{M}_n - \sqrt{(\mathcal{C}_{\mathcal{T},n}-1)(1-\mathcal{M}_n)\mathcal{M}_n} \\
= & \sqrt{1-\mathcal{M}_n}\left\{\sqrt{1-\mathcal{M}_n} - \sqrt{\mathcal{C}_{\mathcal{T},n}\mathcal{M}_n - \mathcal{M}_n}\right\} \\
> & 0,
\end{aligned}$$

where the first equality uses the fact that $M(1-M)$ is decreasing in M for $M \geq 1/2$. By Theorem 1 of Varah (1975), we therefore have

$$\min_{1\leq g\leq N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{-1}||_\infty^{-1} \geq g(\mathcal{C}_{\mathcal{T},n}\mathcal{M}_n; \mathcal{C}_{\mathcal{T},n}),$$

where

$$g(\eta; C) = \frac{C - \eta - \sqrt{(C-1)(C-\eta)\eta}}{C}, \qquad \eta \in [0,1], C \in \mathbb{N}.$$

For every $C \in \mathbb{N}$, $g(\cdot; C)$ continuous and strictly decreasing with $g(1; C) = 0$. Also, $g(\eta; C) \to 1 - \sqrt{\eta}$ as $C \to \infty$. Using these facts, it can be shown that $\inf_{C \in \mathbb{N}} g(\eta; C) > 0$ for every $\eta \in [0, 1)$. As a consequence,

$$\lim_{\delta \downarrow 0} \liminf_{n \to \infty} \mathbb{P}\left[\min_{1 \le g \le N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} > \delta\right] \ge \lim_{\delta \downarrow 0} \liminf_{n \to \infty} \mathbb{P}\left[\mathcal{C}_{\mathcal{T},n}\mathcal{M}_n < 1 - \delta\right].$$

Part (c) now follows from the fact that

$$\mathbb{P}\left[\min_{1 \le g \le N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} > 0\right] \to 1$$

and

$$\frac{1}{\min_{1 \le g \le N_{\mathcal{T},n}} ||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2}} = O_p(1).$$

(d) Suppose $||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} > \sum_{1 \le h \le N_{\mathcal{T},n}, h \ne g} ||\mathbf{M}_n(g,h)||_\infty^2$ holds for every $g = 1, \ldots, N_{\mathcal{T},n}$ and that $\kappa_n = \kappa_n^{\mathtt{CR}} = (\mathbf{M}_n \circledast_n \mathbf{M}_n)^{-1}$, where the inverse exists by Theorem 1 of Feingold and Varga (1962). Then

$$||\kappa_n||_\infty \le \frac{1}{\min_{1 \le g \le N_{\mathcal{T},n}} \left\{ ||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} - \sum_{1 \le h \le N_{\mathcal{T},n}, h \ne g} ||\mathbf{M}_n(g,h)||_\infty^2 \right\}}$$

by Theorem 2 of Varah (1975). Also, $||\mathbf{D}_n(\kappa_n)||_\infty = 0$ by construction. Part (d) now follows from the fact that

$$\mathbb{P}\left[\min_{1 \le g \le N_{\mathcal{T},n}} \left\{ ||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} - \sum_{1 \le h \le N_{\mathcal{T},n}, h \ne g} ||\mathbf{M}_n(g,h)||_\infty^2 \right\} > 0\right] \to 1$$

and

$$\frac{1}{\min_{1 \le g \le N_{\mathcal{T},n}} \left\{ ||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2} - \sum_{1 \le h \le N_{\mathcal{T},n}, h \ne g} ||\mathbf{M}_n(g,h)||_\infty^2 \right\}} = O_p(1).$$

Remark B.5.

- The following assumptions are used in the proof:
  - $\max_{1\leq i\leq n}\mathbb{E}[U_{i,n}^4|\mathcal{X}_n,\mathcal{W}_n]=O_p(1)$
  - $\max_{1\leq i\leq n}\mathbb{E}[\|\mathbf{V}_{i,n}\|^4|\mathcal{W}_n]=O_p(1)$
  - $\chi_n=O(1)$
- If it can be shown that

$$||\mathbf{M}_n(g,g)^{1/2}\otimes\mathbf{M}_n(g,g)^{1/2}-\mathbf{I}_{\#\mathcal{T}_{g,n}}\otimes\mathbf{I}_{\#\mathcal{T}_{g,n}}||_\infty$$

$$\lesssim||\mathbf{M}_n(g,g)\otimes\mathbf{M}_n(g,g)-\mathbf{I}_{\#\mathcal{T}_{g,n}}\otimes\mathbf{I}_{\#\mathcal{T}_{g,n}}||_\infty,$$

  then the condition $\mathcal{C}_{\mathcal{T},n}^3\mathcal{M}_n=o_p(1)$ in part (b) can be weakened to $\mathcal{C}_{\mathcal{T},n}^2\mathcal{M}_n=o_p(1)$.
- Because $||\mathbf{M}_n(g,g)^{-1}||_\infty\leq||\mathbf{M}_n(g,g)^{-1/2}||_\infty^2$, we have

$$\mathbb{P}\left[\min_{1\leq g\leq N_{\mathcal{T},n}}||\mathbf{M}_n(g,g)^{-1/2}||_\infty^{-2}>\delta\right]\leq\mathbb{P}\left[\min_{1\leq g\leq N_{\mathcal{T},n}}||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2}>\delta^2\right]$$

  and therefore

$$\lim_{\delta\downarrow0}\liminf_{n\to\infty}\mathbb{P}\left[\min_{1\leq g\leq N_{\mathcal{T},n}}||\mathbf{M}_n(g,g)^{-1/2}||_\infty^{-2}>\delta\right]$$

$$\leq\lim_{\delta\downarrow0}\liminf_{n\to\infty}\mathbb{P}\left[\min_{1\leq g\leq N_{\mathcal{T},n}}||\mathbf{M}_n(g,g)^{-1}||_\infty^{-2}>\delta\right].$$

  As a consequence, the conditions of part (c) are no stronger than those of part (b).

□

Discussion about the Leave-Out Estimator. Note that

$$\begin{aligned}\hat{\mathbf{\Sigma}}_n^{\mathtt{LO}} &= \operatorname{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'(\mathbf{M}_n(g,g)^{-1}\hat{\mathbf{u}}_n(g)\otimes\mathbf{y}_n(g))\right\}\\ &= \operatorname{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'\kappa_n^{\mathtt{LO}}(g,h)(\hat{\mathbf{u}}_n(h)\otimes\mathbf{y}_n(h))\right\}\\ &\quad \kappa_n^{\mathtt{LO}} = [\operatorname{diag}_n(\mathbf{M}_n\circledast_n\mathbf{I}_n)]^{-1}\end{aligned}$$

Our goal is to show that under certain regularity conditions we have $\hat{\mathbf{\Sigma}}_n^{\mathtt{LO}} = \mathbf{\Sigma}_n + o_p(1)$. Define

$$\tilde{\mathbf{\Sigma}}_n^{\mathtt{LO}} = \operatorname{vec}_d^{-1}\left\{\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'\kappa_n^{\mathtt{LO}}(g,h)[\bar{\mathbf{U}}_n(h)\otimes\mathbf{U}_n(h)]\right\}$$

$$\bar{\mathbf{\Sigma}}_n^{\mathtt{LO}} = \mathbb{E}[\tilde{\mathbf{\Sigma}}_n^{\mathtt{LO}}|\mathcal{X}_n,\mathcal{W}_n]$$

Without loss of generality we assume d=1 like in the proof of other technical lemmas. It suffices to show $\hat{\mathbf{\Sigma}}_n^{\mathtt{LO}} = \tilde{\mathbf{\Sigma}}_n^{\mathtt{LO}} + o_p(1)$, $\tilde{\mathbf{\Sigma}}_n^{\mathtt{LO}} = \bar{\mathbf{\Sigma}}_n^{\mathtt{LO}} + o_p(1)$ and under our choice of $\kappa_n^{\mathtt{LO}}$ we have $\bar{\mathbf{\Sigma}}_n^{\mathtt{LO}} = \mathbf{\Sigma}_n + o_p(1)$.

In order to show $\hat{\mathbf{\Sigma}}_n^{\mathtt{LO}} = \tilde{\mathbf{\Sigma}}_n^{\mathtt{LO}} + o_p(1)$, let us first introduce the decomposition

$$\hat{\mathbf{u}}_n(h) - \bar{\mathbf{U}}_n(h) = \bar{\mathbf{R}}_n(h) - \hat{\mathbf{V}}_n(h)(\hat{\beta}_n - \beta) \qquad \bar{\mathbf{R}}_n(h) = \bar{\mathbf{r}}_n(h) + (\bar{\mathbf{R}}_n(h) - \bar{\mathbf{r}}_n(h))$$

$$\hat{\mathbf{V}}_n(h) = \bar{\mathbf{V}}_n(h) + \bar{\mathbf{Q}}_n(h)$$

Define $\mu_{i,n} = \mathbb{E}[y_{i,n}|\mathcal{X}_n,\mathcal{W}_n]$

$$\boldsymbol{\mu}_n(g) = (\mu_{t_{g,n}(1),n},\ldots,\mu_{t_{g,n}(\#\mathcal{T}_{g,n}),n})', \qquad g = 1,\ldots,N_{\mathcal{T},n},$$

We also have $\mathbf{y}_n(g) = \boldsymbol{\mu}_n(g) + \mathbf{U}_n(g)$.

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_n^{\texttt{LO}} &= \frac{1}{n} \sum_{1 \leq g,h \leq N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(g) \otimes \hat{\mathbf{V}}_n(g))' \kappa_n^{\texttt{LO}}(g,h) \\
&\quad [(\bar{\mathbf{U}}_n(h) + \bar{\mathbf{R}}_n(h) - \hat{\mathbf{V}}_n(h)(\hat{\beta}_n - \beta)) \otimes (\boldsymbol{\mu}_n(h) + \mathbf{U}_n(h))] \\
&= \tilde{\boldsymbol{\Sigma}}_n^{\texttt{LO}} + \frac{1}{n} \sum_{1 \leq g,h \leq N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(g) \otimes \hat{\mathbf{V}}_n(g))' \kappa_n^{\texttt{LO}}(g,h) \\
&\quad [(\bar{\mathbf{R}}_n(h) - \hat{\mathbf{V}}_n(h)(\hat{\beta}_n - \beta)) \otimes \mathbf{U}_n(h)] \\
&+ \frac{1}{n} \sum_{1 \leq g,h \leq N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(g) \otimes \hat{\mathbf{V}}_n(g))' \kappa_n^{\texttt{LO}}(g,h) [\hat{\mathbf{u}}_n(h) \otimes \boldsymbol{\mu}_n(h)]
\end{aligned}$$

To deal with the first term, we will repeatedly using the inequality introduced in the proof of Technical Lemma B.3

$$\begin{aligned}
&\left| \tfrac{1}{4n} \textstyle\sum_{1 \leq g,h \leq N_{\mathcal{T},n}} [\mathbf{x}_n(g) \otimes \mathbf{X}_n(g)] \kappa_n(g,h) [\mathbf{y}_n(g) \otimes \mathbf{Y}_n(g)] \right| \\
&\leq \sqrt{\tfrac{1}{4n} \textstyle\sum^{\kappa,n} \left\{ (x^2_{t_{g,n}(k),n} + x^2_{t_{g,n}(K),n})(y^2_{t_{g,n}(l),n} + y^2_{t_{g,n}(L),n}) \right\}} \\
&\sqrt{\tfrac{1}{4n} \textstyle\sum^{\kappa,n} \left\{ (X^2_{t_{g,n}(k),n} + X^2_{t_{g,n}(K),n})(Y^2_{t_{g,n}(l),n} + Y^2_{t_{g,n}(L),n}) \right\}}
\end{aligned}$$

From the proof in technical lemma B.3, we have already known

$$\frac{1}{4n} \sum^{\kappa,n} \left\{ (\tilde{V}^2_{t_{g,n}(k),n} + \tilde{V}^2_{t_{g,n}(K),n})(\tilde{r}^2_{t_{g,n}(l),n} + \tilde{r}^2_{t_{g,n}(L),n}) \right\} = O_p(\mathcal{C}_{\mathcal{S},n} \mathcal{C}_{\mathcal{T},n} \rho_n)$$

$$\begin{aligned}
&\frac{1}{4n} \sum^{\kappa,n} \left\{ (\tilde{V}^2_{t_{g,n}(k),n} + \tilde{V}^2_{t_{g,n}(K),n})[(\tilde{R}_{t_{g,n}(l),n} - \tilde{r}_{t_{g,n}(l),n})^2 + (\tilde{R}_{t_{g,n}(L),n} - \tilde{r}_{t_{g,n}(L),n})^2] \right\} \\
&\qquad = O_p(\mathcal{C}_{\mathcal{T},n} n(\varrho_n - \rho_n))
\end{aligned}$$

$$\frac{1}{4n} \sum^{\kappa,n} \left\{ (\tilde{Q}^2_{t_{g,n}(k),n} + \tilde{Q}^2_{t_{g,n}(K),n})(\tilde{R}^2_{t_{g,n}(l),n} + \tilde{R}^2_{t_{g,n}(L),n}) \right\} = O_p(\mathcal{C}_{\mathcal{T},n} n \chi_n \varrho_n)$$

$$\frac{(\hat{\beta}_n - \beta)^2}{4n} \sum^{\kappa,n} \left\{ \hat{v}^4_{t_{g,n}(k),n} + \hat{v}^4_{t_{g,n}(K),n} + \hat{v}^4_{t_{g,n}(l),n} + \hat{v}^4_{t_{g,n}(L),n} \right\} = \left(\frac{\mathcal{C}^2_{\mathcal{T},n}}{n^2} \sum_{1 \leq i \leq n} \hat{v}^4_{i,n} \right) O_p(1)$$

It still remains to show the magnitude of

$$\frac{1}{4n}\sum^{\kappa,n}\left\{(\hat{v}^2_{t_{g,n}(k),n}+\hat{v}^2_{t_{g,n}(K),n})(U^2_{t_{g,n}(l),n}+U^2_{t_{g,n}(L),n})\right\}=O_p(\mathcal{C}_{\mathcal{T},n})$$

because

$$\begin{aligned}&\mathbb{E}\left[\tfrac{1}{4n}\textstyle\sum^{\kappa,n}\left\{(\hat{v}^2_{t_{g,n}(k),n}+\hat{v}^2_{t_{g,n}(K),n})(U^2_{t_{g,n}(l),n}+U^2_{t_{g,n}(L),n})\right\}\middle|\,\mathcal{X}_n,\mathcal{W}_n\right]\\&=\tfrac{1}{4n}\textstyle\sum^{\kappa,n}\left\{(\hat{v}^2_{t_{g,n}(k),n}+\hat{v}^2_{t_{g,n}(K),n})\mathbb{E}\left[(U^2_{t_{g,n}(l),n}+U^2_{t_{g,n}(L),n})|\mathcal{X}_n,\mathcal{W}_n\right]\right\}\\&\leq\tfrac{\mathcal{C}_{\mathcal{T},n}\mathcal{C}_{U,n}}{n}\textstyle\sum_{1\leq g\leq N_{\mathcal{T},n}}\sum_{t\in\mathcal{T}_{g,n}}\hat{v}^2_{t,n}=\tfrac{\mathcal{C}_{\mathcal{T},n}\mathcal{C}_{U,n}}{n}\sum_{i=1}^{n}\hat{v}^2_{i,n}=O_p(\mathcal{C}_{\mathcal{T},n})\end{aligned}$$

Combine all the results above, we can show that

$$\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'\kappa_n^{\texttt{LO}}(g,h)[(\bar{\mathbf{R}}_n(h)-\hat{\mathbf{V}}_n(h)(\hat{\beta}_n-\beta))\otimes\mathbf{U}_n(h)]=o_p(1)$$

Then we can deal with the term

$$\begin{aligned}&\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'\kappa_n^{\texttt{LO}}(g,h)[\hat{\mathbf{u}}_n(h)\otimes\boldsymbol{\mu}_n(h)]\\&=\frac{1}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}}\hat{\mathbf{V}}_n(g)'\mathbf{M}_n(g,g)^{-1}\hat{\mathbf{u}}_n(g)\boldsymbol{\mu}_n(g)'\hat{\mathbf{V}}_n(g)\end{aligned}$$

Still we apply the decomposition $\hat{\mathbf{u}}_n(h)-\bar{\mathbf{U}}_n(h)=\bar{\mathbf{R}}_n(h)-\hat{\mathbf{V}}_n(h)(\hat{\beta}_n-\beta)$ and define $\max_{1\leq g\leq\ N_{\mathcal{T},n}}\lambda_{\max}(\mathbf{M}_n(g,g)^{-1})=\mathcal{C}_{\kappa^{LO},n}$

$$
\begin{aligned}
& \left(\tfrac{1}{n}\textstyle\sum_{1\le g\le N_{\mathcal{T},n}} \hat{\mathbf{V}}_n(g)'\mathbf{M}_n(g,g)^{-1}\bar{\mathbf{R}}_n(g)\boldsymbol{\mu}_n(g)'\hat{\mathbf{V}}_n(g)\right)^2 \\
\le\ & \left(\tfrac{1}{n}\textstyle\sum_{1\le g\le N_{\mathcal{T},n}} \{\max_{1\le h\le N_{\mathcal{T},n}} \|\boldsymbol{\mu}_n(h)\|\}\left\|\hat{\mathbf{V}}_n(g)\right\|\hat{\mathbf{V}}_n(g)'\mathbf{M}_n(g,g)^{-1}\bar{\mathbf{R}}_n(g)\right)^2 \\
\le\ & \tfrac{1}{n^2}\{\max_{1\le h\le N_{\mathcal{T},n}} \|\boldsymbol{\mu}_n(h)\|\}^2\left(\textstyle\sum_{1\le g\le N_{\mathcal{T},n}} \left\|\hat{\mathbf{V}}_n(g)\right\|\hat{\mathbf{V}}_n(g)'\mathbf{M}_n(g,g)^{-1}\bar{\mathbf{R}}_n(g)\right)^2 \\
\le\ & \tfrac{\mathcal{C}^2_{\lambda,\kappa^{LO}}}{n^2}\{\max_{1\le h\le N_{\mathcal{T},n}} \|\boldsymbol{\mu}_n(h)\|\}^2\left(\textstyle\sum_{1\le g\le N_{\mathcal{T},n}} \left\|\hat{\mathbf{V}}_n(g)\right\|^2\bar{\mathbf{R}}_n(g)\right)^2 \\
\le\ & \tfrac{\mathcal{C}^2_{\lambda,\kappa^{LO}}}{n^2}\{\max_{1\le h\le N_{\mathcal{T},n}} \|\boldsymbol{\mu}_n(h)\|\}^2\left(\textstyle\sum_{1\le g\le N_{\mathcal{T},n}} \left\|\hat{\mathbf{V}}_n(g)\right\|\hat{\mathbf{V}}_n(g)'\bar{\mathbf{R}}_n(g)\right)^2 \\
\le\ & \tfrac{\mathcal{C}^2_{\lambda,\kappa^{LO}}}{n^2}\{\max_{1\le h\le N_{\mathcal{T},n}} \|\boldsymbol{\mu}_n(h)\|\}^2\textstyle\sum_{1\le g\le N_{\mathcal{T},n}}\left(\left\|\hat{\mathbf{V}}_n(g)\right\|^2\right)^2\sum_{1\le i\le n}\tilde{R}^2_{i,n} \\
\le\ & \mathcal{C}^2_{\lambda,\kappa^{LO}}\mathcal{C}_{\mathcal{T},n}\{\max_{1\le h\le N_{\mathcal{T},n}} \|\boldsymbol{\mu}_n(h)\|\}^2\left(\tfrac{1}{n^2}\textstyle\sum_{i=1}^n \hat{v}^4_{i,n}\right)n\varrho_n
\end{aligned}
$$

We would like to assume that $\mathcal{C}_{\mathcal{T},n}\{\max_{1\le h\le N_{\mathcal{T},n}} \|\boldsymbol{\mu}_n(h)\|\}^2\left(\frac{1}{n^2}\sum_{i=1}^n \hat{v}^4_{i,n}\right)n\varrho_n = o_p(1)$ and $\mathcal{C}_{\lambda,\kappa^{LO}} = O_p(1)$.
Similarly we can have

$$
\begin{aligned}
& \left(\tfrac{1}{n}\textstyle\sum_{1\le g\le N_{\mathcal{T},n}} \hat{\mathbf{V}}_n(g)'\mathbf{M}_n(g,g)^{-1}\hat{\mathbf{V}}_n(g)(\hat{\beta}_n-\beta)\boldsymbol{\mu}_n(g)'\hat{\mathbf{V}}_n(g)\right)^2 \\
\le\ & \tfrac{\mathcal{C}^2_{\lambda,\kappa^{LO}}}{n^2}\{\max_{1\le h\le N_{\mathcal{T},n}} \|\boldsymbol{\mu}_n(h)\|\}^2\textstyle\sum_{1\le g\le N_{\mathcal{T},n}}\left(\left\|\hat{\mathbf{V}}_n(g)\right\|^2\right)^2\sum_{1\le i\le n}\hat{v}^2_{i,n}(\hat{\beta}_n-\beta)^2 \\
\le\ & \mathcal{C}^2_{\lambda,\kappa^{LO}}\mathcal{C}_{\mathcal{T},n}\{\max_{1\le h\le N_{\mathcal{T},n}} \|\boldsymbol{\mu}_n(h)\|\}^2\left(\tfrac{1}{n^2}\textstyle\sum_{i=1}^n \hat{v}^4_{i,n}\right)\left(\tfrac{1}{n}\textstyle\sum_{1\le i\le n}\hat{v}^2_{i,n}\right)n(\hat{\beta}_n-\beta)^2
\end{aligned}
$$

This requires $\mathcal{C}^2_{\mathcal{T},n}\{\max_{1\le h\le N_{\mathcal{T},n}} \|\boldsymbol{\mu}_n(h)\|\}^2\left(\frac{1}{n^2}\sum_{i=1}^n \hat{v}^4_{i,n}\right) = o_p(1)$.
We are left with the term

$$
\frac{1}{n}\sum_{1\le g\le N_{\mathcal{T},n}} \hat{\mathbf{V}}_n(g)'\mathbf{M}_n(g,g)^{-1}\bar{\mathbf{U}}_n(g)\boldsymbol{\mu}_n(g)'\hat{\mathbf{V}}_n(g)
$$

which is conditional mean zero.

$$\begin{aligned}
&\mathbb{E}\left[(\tfrac{1}{n}\textstyle\sum_{g=1}^{N_{\mathcal{T},n}}\hat{\mathbf{V}}_n(g)'\mathbf{M}_n(g,g)^{-1}\bar{\mathbf{U}}_n(g)\boldsymbol{\mu}_n(g)'\hat{\mathbf{V}}_n(g))^2|\mathcal{X}_n,\mathcal{W}_n\right]\\
=\ &\tfrac{1}{n^2}\textstyle\sum_{g=1}^{N_{\mathcal{T},n}}\mathbb{E}\left[(\hat{\mathbf{V}}_n(g)'\mathbf{M}_n(g,g)^{-1}(\sum_{h=1}^{N_{\mathcal{T},n}}\mathbf{M}_n(g,h)\mathbf{U}_n(h))\boldsymbol{\mu}_n(g)'\hat{\mathbf{V}}_n(g))^2|\mathcal{X}_n,\mathcal{W}_n\right]\\
=\ &\tfrac{1}{n^2}\textstyle\sum_{g=1}^{N_{\mathcal{T},n}}\sum_{h=1}^{N_{\mathcal{T},n}}\mathbb{E}\left[(\hat{\mathbf{V}}_n(g)'\mathbf{M}_n(g,g)^{-1}\mathbf{M}_n(g,h)\mathbf{U}_n(h)\boldsymbol{\mu}_n(g)'\hat{\mathbf{V}}_n(g))^2|\mathcal{X}_n,\mathcal{W}_n\right]
\end{aligned}$$

Here we define $\mathcal{C}_{\lambda,U}=\max_{1\leq g\leq N_{\mathcal{T},n}}\lambda_{\max}\mathbb{E}[\mathbf{U}_g\mathbf{U}_g'|\mathcal{X}_n,\mathcal{W}_n]$.
We have that $\sum_{1\leq h\leq N_{\mathcal{T},n}}\mathbf{M}_n(g,h)\mathbf{M}_n(g,h)=\mathbf{M}_n(g,g)$. Apply these two things and similar logic from above process we have

$$\begin{aligned}
&\mathbb{E}\left[(\tfrac{1}{n}\textstyle\sum_{1\leq g\leq N_{\mathcal{T},n}}\hat{\mathbf{V}}_n(g)'\mathbf{M}_n(g,g)^{-1}\bar{\mathbf{U}}_n(g)\boldsymbol{\mu}_n(g)'\hat{\mathbf{V}}_n(g))^2|\mathcal{X}_n,\mathcal{W}_n\right]\\
\leq\ &\tfrac{\mathcal{C}_{\lambda,U}\mathcal{C}_{\lambda,\kappa^{LO}}}{n^2}\{\max_{1\leq h\leq N_{\mathcal{T},n}}\|\boldsymbol{\mu}_n(h)\|\}^2\textstyle\sum_{1\leq g\leq N_{\mathcal{T},n}}\left(\left\|\hat{\mathbf{V}}_n(g)\right\|^2\right)^2\\
\leq\ &\mathcal{C}_{\lambda,U}\mathcal{C}_{\lambda,\kappa^{LO}}\mathcal{C}_{\mathcal{T},n}\{\max_{1\leq h\leq N_{\mathcal{T},n}}\|\boldsymbol{\mu}_n(h)\|\}^2\left(\tfrac{1}{n^2}\textstyle\sum_{i=1}^{n}\hat{v}_{i,n}^4\right)
\end{aligned}$$

We need $\mathcal{C}_{\lambda,U}\mathcal{C}_{\mathcal{T},n}\{\max_{1\leq h\leq N_{\mathcal{T},n}}\|\boldsymbol{\mu}_n(h)\|\}^2\left(\frac{1}{n^2}\sum_{i=1}^{n}\hat{v}_{i,n}^4\right)=o_p(1)$.
Then we have $\hat{\boldsymbol{\Sigma}}_n^{\mathtt{LO}}=\tilde{\boldsymbol{\Sigma}}_n^{\mathtt{LO}}+o_p(1)$.
Next step we will show $\tilde{\boldsymbol{\Sigma}}_n^{\mathtt{LO}}=\bar{\boldsymbol{\Sigma}}_n^{\mathtt{LO}}+o_p(1)$ which means we will focus on $\mathbb{V}[\tilde{\boldsymbol{\Sigma}}_n^{\mathtt{LO}}|\mathcal{X}_n,\mathcal{W}_n]$

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}}_n^{\mathtt{LO}} &= \frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'\kappa_n^{\mathtt{LO}}(g,h)[\bar{\mathbf{U}}_n(h)\otimes\mathbf{U}_n(h)]\\
&= \frac{1}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}}d_{gg,n}+\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n},g<h}(d_{gh,n}+d_{hg,n})
\end{aligned}$$

where

$$d_{gh,n}=\sum_{1\leq k,l\leq N_{\mathcal{T},n}}(\hat{\mathbf{V}}_n(k)\otimes\hat{\mathbf{V}}_n(k))'\kappa_n^{\mathtt{LO}}(k,l)(\mathbf{M}_n(l,g)\otimes\mathbf{I}_n(l,h))[\mathbf{U}_n(g)\otimes\mathbf{U}_n(h)]$$

Use the same logic from proof of technical lemma B.3

$$\begin{aligned}
&\sum_{1\leq g,h\leq N_{\mathcal{T},n}} \mathbb{V}[d_{gh,n}|\mathcal{X}_n,\mathcal{W}_n] \\
&\leq \mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n} \sum_{1\leq l,k,L,K\leq N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(k)\otimes\hat{\mathbf{V}}_n(k))'\kappa_n^{\mathtt{LO}}(k,l)(\mathbf{M}_n(l,L)\otimes\mathbf{I}_n(l,L)) \\
&\quad \kappa_n^{\mathtt{LO}}(L,K)(\hat{\mathbf{V}}_n(K)\otimes\hat{\mathbf{V}}_n(K))' \\
&= \mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n}(\hat{\mathbf{V}}_n\circledast_n\hat{\mathbf{V}}_n)'\kappa_n^{\mathtt{LO}\prime}(\mathbf{M}_n\circledast_n\mathbf{I}_n)\kappa_n^{\mathtt{LO}}(\hat{\mathbf{V}}_n\circledast_n\hat{\mathbf{V}}_n)' \\
&\leq \mathcal{C}_{\kappa,n}^2\mathcal{C}_{\mathcal{T},n}^2\mathcal{C}_{U,n}\|\mathbf{M}_n\circledast_n\mathbf{I}_n\|_\infty\left\|\hat{\mathbf{V}}_n\circledast_n\hat{\mathbf{V}}_n\right\|^2 \\
&\leq \mathcal{C}_{\kappa,n}^2\mathcal{C}_{\mathcal{T},n}^3\mathcal{C}_{U,n}\sum_{1\leq i\leq N_{\mathcal{T},n}}\hat{v}_{i,n}^4 = o_p(n^2)
\end{aligned}$$

where we use

$$\|\mathbf{M}_n\circledast_n\mathbf{I}_n\|_\infty \leq \|\mathbf{M}_n\|_\infty = O_p(1)$$

As a consequence

$$\begin{aligned}
\mathbb{V}\left[\frac{1}{n}\sum_{1\leq g\leq N_{\mathcal{T},n}} d_{gg,n}|\mathcal{X}_n,\mathcal{W}_n\right] &= \frac{1}{n^2}\sum_{1\leq g\leq N_{\mathcal{T},n}}\mathbb{V}\left[d_{gg,n}|\mathcal{X}_n,\mathcal{W}_n\right] \\
&\leq \frac{1}{n^2}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}\mathbb{V}\left[d_{gh,n}|\mathcal{X}_n,\mathcal{W}_n\right] = o_p(1)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{V}\left[\frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n},g<h} d_{gh,n}+d_{hg,n}|\mathcal{X}_n,\mathcal{W}_n\right] &= \frac{1}{n^2}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}^{g<h}\mathbb{V}\left[d_{gh,n}+d_{hg,n}|\mathcal{X}_n,\mathcal{W}_n\right] \\
&\leq \frac{2}{n^2}\sum_{1\leq g,h\leq N_{\mathcal{T},n}}\mathbb{V}\left[d_{gh,n}|\mathcal{X}_n,\mathcal{W}_n\right] = o_p(1)
\end{aligned}$$

implying in particular that $\tilde{\mathbf{\Sigma}}_n^{\mathtt{LO}} = \bar{\mathbf{\Sigma}}_n^{\mathtt{LO}} + o_p(1)$.

Further

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_n^{\mathtt{LO}} - \mathbf{\Sigma}_n &= \frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'\kappa_n^{\mathtt{LO}}(g,h)\mathbb{E}[\bar{\mathbf{U}}_n(h)\otimes\mathbf{U}_n(h)|\mathcal{X}_n,\mathcal{W}_n] \\ &= \frac{1}{n}\sum_{1\leq g,h\leq N_{\mathcal{T},n}} (\hat{\mathbf{V}}_n(g)\otimes\hat{\mathbf{V}}_n(g))'(\kappa_n^{\mathtt{LO}}(g,h)(\mathbf{M}_n\circledast_n\mathbf{I}_n) - \mathbf{I}_{N_{\kappa,n}}) \\ &\quad \mathbb{E}[\mathbf{U}_n(h)\otimes\mathbf{U}_n(h)|\mathcal{X}_n,\mathcal{W}_n]\end{aligned}$$

By construction $\kappa_n^{\mathtt{LO}} = [\mathrm{diag}_n(\mathbf{M}_n\circledast_n\mathbf{I}_n)]^{-1} = (\mathbf{M}_n\circledast_n\mathbf{I}_n)^{-1}$.

Remark B.6. Besides the assumptions in technical lemma B.3, some extra assumptions are used in the proof

- $\mathcal{C}_{\lambda,\kappa^{LO}} = O_p(1)$.
- $\mathcal{C}_{\mathcal{T},n}\{\max_{1\leq h\leq N_{\mathcal{T},n}}\|\boldsymbol{\mu}_n(h)\|\}^2\left(\frac{1}{n^2}\sum_{i=1}^n\|\hat{\mathbf{v}}_{i,n}\|^4\right)n\varrho_n = o_p(1)$.
- $\mathcal{C}^2_{\mathcal{T},n}\{\max_{1\leq h\leq N_{\mathcal{T},n}}\|\boldsymbol{\mu}_n(h)\|\}^2\left(\frac{1}{n^2}\sum_{i=1}^n\|\hat{\mathbf{v}}_{i,n}\|^4\right) = o_p(1)$.
- $\mathcal{C}_{\lambda,U}\mathcal{C}_{\mathcal{T},n}\{\max_{1\leq h\leq N_{\mathcal{T},n}}\|\boldsymbol{\mu}_n(h)\|\}^2\left(\frac{1}{n^2}\sum_{i=1}^n\|\hat{\mathbf{v}}_{i,n}\|^4\right) = o_p(1)$.
- $(\mathbf{M}_n\circledast_n\mathbf{I}_n)^{-1}$ exists.

□

APPENDIX C

Proofs for Chapter IV

Proof of Lemma IV.1. According to (IV.5), the second moment condition is

$$\begin{aligned}\mathbb{E}(y) &= \mathbb{E}(X(\bar{s},s)+u) = X(\bar{s},s) \\ &= \frac{1}{2\lambda_3}\left[\bar{v}+\frac{\rho\sigma_v}{\sigma_s}(s-\bar{s})-\lambda_1-\lambda_2 s\right].\end{aligned}$$

Computation of (IV.5) and the requirement that X is affine imply that we can focus on the case with $\lambda_3 > 0$. Define an unbiased estimator for $\bar{s}$:

$$s_y := \frac{\sigma_s}{\rho\sigma_v}\left(\bar{v}-\lambda_1+\left(\frac{\rho\sigma_v}{\sigma_s}-\lambda_2\right)s-2\lambda_3 y\right).$$

It is unbiased because

$$\begin{aligned}s_y &= \frac{\sigma_s}{\rho\sigma_v}\left(\bar{v}-\lambda_1+\left(\frac{\rho\sigma_v}{\sigma_s}-\lambda_2\right)s-2\lambda_3(X(\bar{s},s)+u)\right) \\ &= \bar{s}-\frac{2\lambda_3\sigma_s}{\rho\sigma_v}u.\end{aligned}$$

Since $\bar{s}-s_y$ is equal to some nonzero constant times $X(\bar{s},s)-y$,

$$\mathbb{E}[s_y] = \bar{s}$$

is an equivalent way to write down the second moment condition $\mathbb{E}(y) = X(\bar{s}, s)$. It may affect the optimal weighting matrix, but not the optimal GMM estimator.

Next, we solve the GMM minimization problem (see (IV.6)) with moment conditions

$$\mathbb{E}[s_y] = \bar{s} \text{ and } \mathbb{E}[s] = \bar{s}.$$

It is useful to write s as $s = \bar{s} + \varepsilon_s$ in which $\varepsilon_s \sim \mathcal{N}(0, \sigma_s^2)$, and $s_y = \bar{s} + \varepsilon_{s_y}$ in which $\varepsilon_{s_y} = -\frac{2\lambda_3\sigma_s}{\rho\sigma_v}u$. Since $\begin{bmatrix} s \\ s_y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \bar{s} \\ \bar{s} \end{bmatrix}, \begin{bmatrix} \sigma_s^2 & 0 \\ 0 & \left(\frac{2\lambda_3\sigma_s\sigma_u}{\rho\sigma_v}\right)^2 \end{bmatrix}\right)$, standard econometric arguments imply that the optimal GMM weighting matrix is

$$\left(\mathbb{E}\left[\begin{bmatrix} \bar{s} - s \\ \bar{s} - s_y \end{bmatrix} \times \begin{bmatrix} \bar{s} - s \\ \bar{s} - s_y \end{bmatrix}'\right]\right)^{-1} = \begin{bmatrix} \frac{1}{\sigma_s^2} & 0 \\ 0 & \left(\frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\right)^2 \end{bmatrix}.$$

Therefore, the GMM minimization problem becomes the problem of finding the estimator for $\bar{s}$ that minimizes $\frac{1}{\sigma_s^2}(\bar{s} - s)^2 + \left(\frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\right)^2 (\bar{s} - s_y)^2$, which is equivalent to the following linear regression problem:

$$\begin{bmatrix} \frac{1}{\sigma_s}s \\ \frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}s_y \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_s} \\ \frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u} \end{bmatrix} \times \bar{s} + \begin{bmatrix} \frac{1}{\sigma_s}\varepsilon_s \\ \frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\varepsilon_{s_y} \end{bmatrix}. \tag{C.1}$$

Since $\frac{1}{\sigma_s}\varepsilon_s$ and $\frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\varepsilon_{s_y}$ both follow the standard normal distribution and we have $\mathrm{Cov}\left(\frac{1}{\sigma_s}\varepsilon_s, \frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\varepsilon_{s_y}\right) = 0$, by the Gauss–Markov theorem, we conclude that the OLS estimator of (C.1), and hence the optimal GMM estimator,

$$\hat{s}_{GMM}(s, y) = \hat{s}_{Transformed\_OLS} = \frac{\mathrm{Var}(\varepsilon_s)^{-1} \times s + \mathrm{Var}(\varepsilon_{s_y})^{-1} \times s_y}{\mathrm{Var}(\varepsilon_s)^{-1} + \mathrm{Var}(\varepsilon_{s_y})^{-1}}, \tag{C.2}$$

is the unique BLUE of $\bar{s}$ among all linear unbiased estimators of $\bar{s}$. □

Proof of Proposition IV.2. Given an arbitrary affine trading strategy $X(\bar{s}, s) = \alpha_1 + \alpha_2 s + \alpha_3\bar{s}$, we want to find the pricing strategy $P^r(s, y)$ that solves the robustness

problem

$$\min_{\tilde{P}(s,y) \text{ is affine}} \max_{\tilde{s}\in\mathbb{R}} \mathbb{E}_{\tilde{s}}[c(|v - \tilde{P}(s,y)|)].$$

Assuming that $\tilde{P}(s,y) = \lambda_1 + \lambda_2 s + \lambda_3 y$ and plugging in $y = X(\bar{s}, s) + u$, the robustness problem becomes

$$\min_{\lambda_1,\lambda_2,\lambda_3} \max_{\tilde{s}\in\mathbb{R}} \mathbb{E}_{\tilde{s}}[c(|v - \lambda_1 - \lambda_2 s - \lambda_3(\alpha_1 + \alpha_2 s + \alpha_3 \tilde{s} + u)|)].$$

Notice that $v - \lambda_1 - \lambda_2 s - \lambda_3(\alpha_1 + \alpha_2 s + \alpha_3 \tilde{s} + u)$ is a normal random variable whose mean is $\mu = \bar{v} - \lambda_1 - \lambda_3\alpha_1 - (\lambda_2 + \lambda_3\alpha_2 + \lambda_3\alpha_3)\tilde{s}$ and variance is $\sigma^2 = \sigma_v^2 + (\lambda_2 + \lambda_3\alpha_2)^2\sigma_s^2 + \lambda_3^2\sigma_u^2 - 2(\lambda_2 + \lambda_3\alpha_2)\rho\sigma_v\sigma_s$.

Note that the mean μ is affine in $\tilde{s}$, the variance σ^2 is independent of $\tilde{s}$, and the loss function c is unbounded. Therefore, as long as in μ the coefficient of $\tilde{s}$ is nonzero, the worst-case expected loss will always go to infinity after maximizing over $\tilde{s}$. This can be formalized via the following claim:

$$\text{For a random variable} \quad Z \sim \mathcal{N}(\mu, \sigma^2),\ \mathbb{E}[c(|Z|)] \to +\infty \quad \text{if} \quad \mu \to \infty.$$

To verify this claim, first, without loss of generality, let $\mu > 0$. Then,

$$\begin{aligned}
\mathbb{E}[c(|Z|)] &= \int_{-\infty}^{+\infty} c(|Z|)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(Z-\mu)^2}{2\sigma^2}}dZ = \int_{-\infty}^{+\infty} c(|\mu + \sigma t|)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \\
&\geq \int_{-\frac{\mu}{\sigma}}^{+\infty} c(\mu + \sigma t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \geq \int_{0}^{+\infty} c(\mu + \sigma t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \\
&\geq \int_{0}^{+\infty} c(\mu)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt.
\end{aligned}$$

The last term, $\int_0^{+\infty} c(\mu)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$, will go to $+\infty$ if $\mu \to +\infty$.

Based on this observation, the coefficient of $\tilde{s}$ in μ must be zero; that is, $\lambda_2 + \lambda_3\alpha_2 + \lambda_3\alpha_3 = 0$. Since $\tilde{s}$ becomes irrelevant, the market maker's robustness problem

can be rewritten as

$$\min_{\lambda_1,\lambda_3} \mathbb{E}[c(|Z|)], \tag{C.3}$$

in which

$$Z \sim \mathcal{N}(\bar{v} - \lambda_1 - \lambda_3\alpha_1, \sigma_v^2 + (\lambda_2 + \lambda_3\alpha_2)^2\sigma_s^2 + \lambda_3^2\sigma_u^2 - 2(\lambda_2 + \lambda_3\alpha_2)\rho\sigma_v\sigma_s)$$

and $\lambda_2 = -\lambda_3\alpha_2 - \lambda_3\alpha_3$. Taking $\lambda_2 = -\lambda_3\alpha_2 - \lambda_3\alpha_3$ into account,

$$Z \sim \mathcal{N}(\bar{v} - \lambda_1 - \lambda_3\alpha_1, \sigma_v^2 + (\lambda_3\alpha_3)^2\sigma_s^2 + \lambda_3^2\sigma_u^2 + 2\lambda_3\alpha_3\rho\sigma_v\sigma_s).$$

To solve this problem, let us first analyze the derivatives of $c(|Z|)$ as a function of Z's mean μ and standard deviation σ. Again, without loss of generality, let $\mu \geq 0$. Let

$$\begin{aligned} f(\mu,\sigma) \quad &: \quad = \mathbb{E}[c(|Z|)] = \int_{-\infty}^{+\infty} c(|Z|)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(Z-\mu)^2}{2\sigma^2}}dZ \\ &= \int_{-\infty}^{+\infty} c(|\mu+\sigma t|)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \\ &= \int_{-\infty}^{-\frac{\mu}{\sigma}} c(-\mu-\sigma t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt + \int_{-\frac{\mu}{\sigma}}^{+\infty} c(\mu+\sigma t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt. \end{aligned}$$

Since $c(0) = 0$,

$$\frac{\partial f}{\partial \mu} = -\int_{-\infty}^{-\frac{\mu}{\sigma}} c'(-\mu-\sigma t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt + \int_{-\frac{\mu}{\sigma}}^{+\infty} c'(\mu+\sigma t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt.$$

In particular, regardless of the value of σ,

$$\left.\frac{\partial f}{\partial \mu}\right|_{\mu=0} = 0.$$

Because $c'' \geq 0$ and $c'' = 0 \Rightarrow c' > 0$,

$$\begin{aligned}\frac{\partial^2 f}{\partial \mu^2} &= \int_{-\infty}^{-\frac{\mu}{\sigma}} c''(-\mu - \sigma t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt + \int_{-\frac{\mu}{\sigma}}^{+\infty} c''(\mu + \sigma t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt \\ &\quad + \frac{2}{\sigma}c'(0)\frac{1}{\sqrt{2\pi}}e^{-\frac{\mu^2}{2\sigma^2}} \\ &> 0.\end{aligned}$$

Therefore, fixing an arbitrary $\sigma^2 > 0$, $\mu = 0$ is the unique minimizer of f. Similar analysis applies to the situation with $\mu \leq 0$.

Next, we show that the solution of (C.3) must have $\mu = 0$. Suppose the solution of (C.3), $\lambda_1^\star$, $\lambda_2^\star$, and $\lambda_3^\star$, implies that the mean of Z is $\mu^\star \neq 0$ and the standard deviation of Z is $\sigma^\star$. Notice that $\lambda_1^\star$ does not appear in $\sigma^\star$. We can replace $\lambda_1^\star$ with $\tilde{\lambda}_1^\star = \bar{v} - \lambda_3^\star \alpha_1$; after this change, the mean of Z becomes zero but $\sigma^\star$ is unaffected. Therefore, in the solution of (C.3), the mean of Z must be zero.

When μ is zero,

$$f(0,\sigma) = \int_{-\infty}^{0} c(-\sigma t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt + \int_{0}^{+\infty} c(\sigma t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$$

and

$$\left.\frac{\partial f}{\partial \sigma}\right|_{\mu=0} = \int_{-\infty}^{0} c'(-\sigma t)(-t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt + \int_{0}^{+\infty} c'(\sigma t)t\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt > 0.$$

This means that to solve (C.3), we only need to find λ_1 and λ_3 that minimize σ and ensure that μ is zero; that is, we can rewrite the market maker's robustness problem one more time:

$$\min_{\lambda_3} \sigma_v^2 + (\lambda_3\alpha_3)^2\sigma_s^2 + \lambda_3^2\sigma_u^2 + 2\lambda_3\alpha_3\rho\sigma_v\sigma_s,$$

with $\bar{v} = \lambda_1 + \lambda_3\alpha_1$ and $\lambda_2 + \lambda_3\alpha_2 + \lambda_3\alpha_3 = 0$. Solving the minimization problem above, we have $\lambda_3 = -\frac{\alpha_3\rho\sigma_v\sigma_s}{\sigma_u^2 + \alpha_3^2\sigma_s^2}$.

Note that $y = \alpha_1 + \alpha_2 s + \alpha_3\bar{s} + u$. If $\alpha_3 \neq 0$, we can follow the proof of Lemma

IV.1 and define $s_y = \frac{y-\alpha_1-\alpha_2 s}{\alpha_3} = \bar{s} + \frac{u}{\alpha_3}$. According to (C.2),

$$\hat{s}(s,y) = \frac{\frac{1}{\sigma_s^2}s + \frac{\alpha_3^2}{\sigma_u^2}s_y}{\frac{1}{\sigma_s^2} + \frac{\alpha_3^2}{\sigma_u^2}}.$$

If $\alpha_3 = 0$, the optimal estimator of $\bar{s}$ is simply s. Hence, with an abuse of notation, we can still write

$$\hat{s}(s,y) = \frac{\frac{1}{\sigma_s^2}s + \frac{\alpha_3^2}{\sigma_u^2}s_y}{\frac{1}{\sigma_s^2} + \frac{\alpha_3^2}{\sigma_u^2}}.$$

Since

$$s - \hat{s}(s,y) = s - \frac{\frac{1}{\sigma_s^2}s + \frac{\alpha_3^2}{\sigma_u^2}s_y}{\frac{1}{\sigma_s^2} + \frac{\alpha_3^2}{\sigma_u^2}} = \frac{\frac{\alpha_3^2}{\sigma_u^2}}{\frac{1}{\sigma_s^2} + \frac{\alpha_3^2}{\sigma_u^2}}(s - s_y) = \frac{\alpha_3^2\sigma_s^2}{\sigma_u^2 + \alpha_3^2\sigma_s^2}(s - s_y),$$

the pricing strategy $P^r(s,y)$ must satisfy

$$\begin{aligned} P^r(s,y) &= \lambda_1 + \lambda_2 s + \lambda_3 y = \lambda_1 + \lambda_2 s + \lambda_3(\alpha_3 s_y + \alpha_1 + \alpha_2 s) \\ &= \lambda_1 + \lambda_3\alpha_1 + \lambda_3\alpha_3(s_y - s) = \bar{v} + \frac{\alpha_3^2\rho\sigma_v\sigma_s}{\sigma_u^2 + \alpha_3^2\sigma_s^2}(s - s_y) \\ &= \bar{v} + \frac{\rho\sigma_v}{\sigma_s}\frac{\alpha_3^2\sigma_s^2}{\sigma_u^2 + \alpha_3^2\sigma_s^2}(s - s_y) = \bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \hat{s}(s,y)) \\ &= \mathbb{E}_{\hat{s}(s,y)}[v|s,y]. \end{aligned}$$

□

Generalization to the Elliptical Distribution. Here we describe how we could extend our results to the elliptical distribution. We focus on the following type of elliptical distributions that generalizes the normal distribution: For a vector of (elliptically distributed) random variables Z, its density function is equal to $k \cdot g((Z-\mu_Z)'\Sigma_Z^{-1}(Z-\mu_Z))$ for some normalizing constant k and some function $g : [0,+\infty) \to [0,+\infty)$, in which we implicitly require that the mean and the covariance matrix of Z, μ_Z and

Σ_Z, exist.[1] This type of distribution is used by Ball (2020) recently to generalize the normal distribution assumption. Now, replace all normal distributions in our setup with such elliptical distributions. We focus on Proposition IV.2 below, but most of our other results can also be generalized.

Proposition IV.2 continues to hold under the elliptical distribution, mainly because of three reasons. First, the elliptical distribution is closed under linear transformations; that is, linear tranformations of elliptically distributed random variables are still elliptical. Second, thanks to the form of the density function of the elliptical distribution, similar to the proof of Proposition IV.2, we can again establish that (i) regardless of the variance of $\tilde{P}(s,y)-v$, $\frac{\partial f}{\partial \mu}\Big|_{\mu=0}=0$ and $\frac{\partial^2 f}{\partial \mu^2}>0$, and (ii) given $\mu=0$, the expected loss is increasing in the variance of $\tilde{P}(s,y)-v$. Last, the elliptical distribution's conditional expectation satisfies $\mathbb{E}_{\hat{s}(s,y)}[v|s]=\bar{v}+\frac{\rho\sigma_v}{\sigma_s}(s-\hat{s}(s,y))$ (see Lemma 1 of Ball (2020)). This property allows us to show that the conditional expectation in the two-step learning procedure and the solution to the robustness problem must coincide. □

Proof of Lemma IV.2. Following the proof of Lemma IV.1, we already know that $\begin{bmatrix} s \\ s_y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \bar{s} \\ \bar{s} \end{bmatrix}, \begin{bmatrix} \sigma_s^2 & 0 \\ 0 & \left(\frac{2\lambda_3\sigma_s\sigma_u}{\rho\sigma_v}\right)^2 \end{bmatrix}\right)$. The maximum likelihood estimator of $\bar{s}$, $\hat{s}_{ML}$, maximizes

$$L=\frac{1}{\sqrt{2\pi}}e^{-\frac{(s-\hat{s}_{ML}(s,y))^2}{2\sigma_s^2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{(s_y-\hat{s}_{ML}(s,y))^2}{2\left(\frac{2\lambda_3\sigma_s\sigma_u}{\rho\sigma_v}\right)^2}},$$

or equivalently,

$$l=\log L=-\log(2\pi)-\frac{(s-\hat{s}_{ML}(s,y))^2}{2\sigma_s^2}-\frac{(s_y-\hat{s}_{ML}(s,y))^2}{2\left(\frac{2\lambda_3\sigma_s\sigma_u}{\rho\sigma_v}\right)^2}.$$

[1] A general elliptical distribution does not require, for example, that the mean vector exist.

Maximizing the log-likelihood function l, we find that

$$\frac{(s - \hat{s}_{ML}(s,y))}{\sigma_s^2} + \frac{(s_y - \hat{s}_{ML}(s,y))}{\left(\frac{2\lambda_3\sigma_s\sigma_u}{\rho\sigma_v}\right)^2} = 0,$$

and hence

$$\hat{s}_{ML}(s,y) = \frac{\frac{1}{\sigma_s^2} \times s + \frac{1}{\left(\frac{2\lambda_3\sigma_s\sigma_u}{\rho\sigma_v}\right)^2} \times s_y}{\frac{1}{\sigma_s^2} + \frac{1}{\left(\frac{2\lambda_3\sigma_s\sigma_u}{\rho\sigma_v}\right)^2}} = \frac{\mathrm{Var}(\varepsilon_s)^{-1} \times s + \mathrm{Var}(\varepsilon_{s_y})^{-1} \times s_y}{\mathrm{Var}(\varepsilon_s)^{-1} + \mathrm{Var}(\varepsilon_{s_y})^{-1}} = \hat{s}_{GMM}(s,y).$$

□

Proof of Theorem IV.1. Let $P_{\hat{s}(s,y)}(s,y) = \lambda_1 + \lambda_2 s + \lambda_3 y$. From Lemma IV.1, we know that

$$\hat{s}(s,y) = \frac{\mathrm{Var}(\varepsilon_s)^{-1} \times s + \mathrm{Var}(\varepsilon_{s_y})^{-1} \times s_y}{\mathrm{Var}(\varepsilon_s)^{-1} + \mathrm{Var}(\varepsilon_{s_y})^{-1}}. \tag{C.4}$$

Plugging in $\mathrm{Var}(\varepsilon_{s_y}) = \left(\frac{2\lambda_3\sigma_s}{\rho\sigma_v}\right)^2 \sigma_u^2$ and $\mathrm{Var}(\varepsilon_s) = \sigma_s^2$, we have

$$\hat{s}(s,y) = \frac{\frac{1}{\sigma_s^2} \times s + \left(\frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\right)^2 \times s_y}{\frac{1}{\sigma_s^2} + \left(\frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\right)^2}.$$

Recall that $s_y = \frac{\sigma_s}{\rho\sigma_v}\left(\bar{v} - \lambda_1 + (\frac{\rho\sigma_v}{\sigma_s} - \lambda_2)s - 2\lambda_3 y\right)$. Definition IV.5 requires that

$$\begin{aligned}
\mathbb{E}_{\hat{s}(s,y)}[v|s,y] &= \bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \hat{s}(s,y)) \\
&= \bar{v} + \frac{\rho\sigma_v}{\sigma_s}\left[s - \frac{\frac{1}{\sigma_s^2} \times s + \left(\frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\right)^2 \times s_y}{\frac{1}{\sigma_s^2} + \left(\frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\right)^2}\right] \\
&= \left[\bar{v} - \frac{\rho\sigma_v}{\sigma_s}\frac{\left(\frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\right)^2 \frac{\sigma_s}{\rho\sigma_v}(\bar{v} - \lambda_1)}{\frac{1}{\sigma_s^2} + \left(\frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\right)^2}\right] + \frac{\rho\sigma_v}{\sigma_s}\frac{\left(\frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\right)^2 \frac{\sigma_s}{\rho\sigma_v}}{\frac{1}{\sigma_s^2} + \left(\frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\right)^2}\lambda_2 s \\
&\quad + 2\frac{\rho\sigma_v}{\sigma_s}\frac{\left(\frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\right)^2 \frac{\sigma_s}{\rho\sigma_v}}{\frac{1}{\sigma_s^2} + \left(\frac{\rho\sigma_v}{2\lambda_3\sigma_s\sigma_u}\right)^2}\lambda_3 y \\
&= P_{\hat{s}(s,y)}(s,y) = \lambda_1 + \lambda_2 s + \lambda_3 y.
\end{aligned}$$

Matching the coefficients, we find $\lambda_1 = \bar{v}$, $\lambda_2 = 0$, and $\lambda_3 = \frac{\rho\sigma_v}{2\sigma_u}$. Therefore,

$$\hat{s}(s,y) = s - \frac{\sigma_s}{2\sigma_u}y$$

and

$$P_{\hat{s}(s,y)}(s,y) = \bar{v} + \frac{\rho\sigma_v}{2\sigma_u}y.$$

Given $\hat{s}(s,y)$ and $P_{\hat{s}(s,y)}(s,y)$, $X(\bar{s},s)$ maximizes $\mathbb{E}_{\bar{s}}[\pi|s]$. Suppose the probabilistically informed trader's order is x. Then,

$$\begin{aligned}
\mathbb{E}_{\bar{s}}[\pi|s] &= \mathbb{E}[(v - P_{\hat{s}(s,y)}(s,y))x|s] \\
&= \mathbb{E}\left[\left(v - \bar{v} - \frac{\rho\sigma_v}{2\sigma_u}(x+u)\right)x \middle| s\right].
\end{aligned}$$

The solution to the maximization of $\mathbb{E}_{\bar{s}}[\pi|s]$ yields

$$X(\bar{s},s) = \frac{\bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}) - \bar{v}}{2\frac{\rho\sigma_v}{2\sigma_u}} = \frac{E_{\bar{s}}[v|s] - \bar{v}}{2\frac{\rho\sigma_v}{2\sigma_u}} = \frac{\sigma_u}{\sigma_s}(s - \bar{s}).$$

This completes the proof of the theorem. □

Proof of Theorem IV.2. Consider arbitrary affine trading strategies

$$X_m(\bar{s}, s, p_1, \ldots, p_{m-1}) = \left(\alpha_m + \beta_{1,m} s - \beta_{2,m} \bar{s} + \sum_{i=1}^{m-1} \gamma_{i,m} p_i\right) \Delta t_m$$

for each $m \in \{1, \ldots, N\}$. Take any $n \in \{1, \ldots, N\}$, and consider arbitrary affine pricing strategies

$$P_m^r(s, y_1, \ldots, y_m) = \lambda_{m,1} + \lambda_{m,2} s + \sum_{i=1}^{m} \lambda_{m,i+2} y_i$$

for each $m < n$.

First, suppose $\beta_{2,m} \neq 0$ for every $m \in \{1, \ldots, N\}$. We can define

$$s_{y,m} = \frac{y_m - \left(\alpha_m + \beta_{1,m} s + \sum_{i=1}^{m-1} \gamma_{i,m} p_i\right) \Delta t_m}{-\beta_{2,m} \Delta t_m} = \bar{s} - \frac{u_m}{\beta_{2,m} \Delta t_m}.$$

Equivalently,

$$y_m = \left(\alpha_m + \beta_{1,m} s - \beta_{2,m} s_{y,m} + \sum_{i=1}^{m-1} \gamma_{i,m} p_i\right) \Delta t_m.$$

Note that $s_{y,m}$ is an affine function of y_m, s, and all past prices $p_1, \ldots, p_{m-1}$, and y_m is an affine function of $s_{y,m}$, s, and $p_1, \ldots, p_{m-1}$.

We claim that every function P_n^r that is affine in $s, y_1, \ldots, y_n$ can be written as an affine function in $s, s_{y,1}, \ldots, s_{y,n}$ uniquely, and vice versa. With an abuse of notation, when we think of P_n^r as a function of $s, y_1, \ldots, y_n$, we write $P_n^r(s, y_1, \ldots, y_n)$, and when we think of it as a function of $s, s_{y,1}, \ldots, s_{y,n}$, we write $P_n^r(s, s_{y,1}, \ldots, s_{y,n})$. We prove this claim by induction. If $n = 1$, this is straightforward. Suppose $n > 1$ and the claim holds for every $m < n$. Without loss of generality, let

$$P_m^r(s, s_{y,1}, \ldots, s_{y,m}) = \eta_{m,1} + \eta_{m,2} s + \eta_{m,3} s_{y,1} + \cdots + \eta_{m,m+2} s_{y,m}$$

for every $m < n$. Take a function P_n^r that is affine in $s, y_1, y_2, \ldots, y_n$:

$$P_n^r(s, y_1, \ldots, y_n) = \lambda_{n,1} + \lambda_{n,2}s + \sum_{i=1}^{n} \lambda_{n,i+2} y_i = \begin{bmatrix} \lambda_{n,1} \\ \lambda_{n,2} \\ \lambda_{n,3} \\ \vdots \\ \lambda_{n,n+2} \end{bmatrix}' \begin{bmatrix} 1 \\ s \\ y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Because

$$\begin{aligned} y_m &= (\alpha_m + \beta_{1,m}s - \beta_{2,m}s_{y,m} + \sum_{i=1}^{m-1} \gamma_{i,m} p_i)\Delta t_m \\ &= (\alpha_m + \beta_{1,m}s - \beta_{2,m}s_{y,m} + \sum_{i=1}^{m-1} \gamma_{i,m} P_i^r(s, s_{y,1}, \ldots, s_{y,i}))\Delta t_m, \end{aligned}$$

we can convert $P_n^r(s, y_1, \ldots, y_n)$ into an affine function $P_n^r(s, s_{y,1}, \ldots, s_{y,n})$ uniquely, as follows:

$$\begin{aligned} P_n^r(s, y_1, \ldots, y_n) &= \lambda_{n,1} + \lambda_{n,2}s + \lambda_{n,3}(\alpha_1 + \beta_{1,1}s - \beta_{2,1}s_{y,1})\Delta t_1 + \ldots \\ &+ \lambda_{n,n+2} \left(\alpha_n + \beta_{1,n}s - \beta_{2,n}s_{y,n} + \sum_{i=1}^{n-1} \gamma_{i,n} P_i^r(s, s_{y,1}, \ldots, s_{y,i}) \right) \Delta t_n \\ &= \begin{bmatrix} \lambda_{n,1} \\ \lambda_{n,2} \\ \lambda_{n,3} \\ \vdots \\ \lambda_{n,n+2} \end{bmatrix}' \Omega \begin{bmatrix} 1 \\ s \\ s_{y,1} \\ \vdots \\ s_{y,n} \end{bmatrix}, \end{aligned}$$

for some $(n+2) \times (n+2)$ matrix Ω. Moreover, this matrix is full-rank, because it is lower triangular and the numbers along its diagonal are $1, 1, \beta_{2,1}\Delta t_1, \ldots, \beta_{2,n}\Delta t_n$, none of which is zero. The fact that the matrix Ω is full-rank also implies that every

$P_n^r(s, s_{y,1}, \ldots, s_{y,n})$ can be rewritten as some $P_n^r(s, y_1, \ldots, y_n)$ uniquely. To see this,

$$P_n^r(s, s_{y,1}, \ldots, s_{y,n}) = \begin{bmatrix} \eta_{n,1} \\ \eta_{n,2} \\ \eta_{n,3} \\ \vdots \\ \eta_{n,n+2} \end{bmatrix}' \begin{bmatrix} 1 \\ s \\ s_{y,1} \\ \vdots \\ s_{y,n} \end{bmatrix} = \begin{bmatrix} \eta_{n,1} \\ \eta_{n,2} \\ \eta_{n,3} \\ \vdots \\ \eta_{n,n+2} \end{bmatrix}' \Omega^{-1} \begin{bmatrix} 1 \\ s \\ y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Thus, without loss of generality, we write pricing strategies in the form of

$$P_m^r(s, s_{y,1}, \ldots, s_{y,m}) = \eta_{m,1} + \eta_{m,2}s + \eta_{m,3}s_{y,1} + \cdots + \eta_{m,m+2}s_{y,m}$$

for every $m \in \{1, \ldots, N\}$.

We need to solve

$$\min_{\tilde{P}_n, \ldots, \tilde{P}_N \text{ are affine}} \max_{\tilde{s} \in \mathbb{R}} \mathbb{E}_{\tilde{s}} \left[\sum_{l=n}^{N} c(|\tilde{P}_l(s, s_{y,1}, \ldots, s_{y,l}) - v|) \right] \tag{C.5}$$

using backward induction. To do so, we start from the last date: The pricing strategy at t_N solves its robustness problem:

$$\min_{\eta_{N,1}, \ldots, \eta_{N,N+2}} \max_{\tilde{s} \in \mathbb{R}} \mathbb{E}_{\tilde{s}}[c(|v - \eta_{N,1} - \eta_{N,2}s - \eta_{N,3}s_{y,1} - \cdots - \eta_{N,N+2}s_{y,N}|)]. \tag{C.6}$$

Note that $v - \eta_{N,1} - \eta_{N,2}s - \eta_{N,3}s_{y,1} - \cdots - \eta_{N,N+2}s_{y,N}$ is a normal random variable with mean $\mu = \bar{v} - \eta_{N,1} - (\eta_{N,2} + \cdots + \eta_{N,N+2})\tilde{s}$ and variance

$$\sigma^2 = \sigma_v^2 + \eta_{N,2}^2\sigma_s^2 - 2\eta_{N,2}\rho\sigma_v\sigma_s + \left(\frac{\eta_{N,3}^2}{\beta_{2,1}^2 \Delta t_1} + \cdots + \frac{\eta_{N,N+2}^2}{\beta_{2,N}^2 \Delta t_N} \right) \sigma_u^2.$$

To solve (C.6), following similar arguments used in the proof of Proposition IV.2, the coefficient of $\tilde{s}$ in μ must be zero, μ must be zero, and σ^2 is minimized given the previous two requirements. In other words, we have $\eta_{N,1} = \bar{v}$ and $\eta_{N,2} = -\eta_{N,3} - \cdots - \eta_{N,N+2}$. Plug $\eta_{N,2} = -\eta_{N,3} - \cdots - \eta_{N,N+2}$ into the above equation for σ^2. Then,

to minimize σ^2, we take partial derivatives of σ^2 with respect to $\eta_{N,3},\ldots,\eta_{N,N+2}$. We find that for any $3 \le m \le N+2$,

$$(\eta_{N,3}+\cdots+\eta_{N,N+2})\sigma_s^2+\rho\sigma_v\sigma_s+\frac{\eta_{N,m}\sigma_u^2}{\beta_{2,m-2}^2\Delta t_{m-2}}=0,$$

which implies that

$$\eta_{N,m}=-\frac{\rho\sigma_v\sigma_s\beta_{2,m-2}^2\Delta t_{m-2}}{(\beta_{2,1}^2\Delta t_1+\cdots+\beta_{2,N}^2\Delta t_N)\sigma_s^2+\sigma_u^2}=-\frac{\rho\sigma_v}{\sigma_s}\frac{\frac{\beta_{2,m-2}^2\Delta t_{m-2}}{\sigma_u^2}}{\frac{1}{\sigma_s^2}+\frac{\beta_{2,1}^2\Delta t_1+\cdots+\beta_{2,N}^2\Delta t_N}{\sigma_u^2}}.$$

Then, we know that

$$\begin{aligned}P_N^r(s,s_{y,1},\ldots,s_{y,N}) &= \bar{v}+\frac{\rho\sigma_v}{\sigma_s}\frac{\sum_{i=1}^N\frac{\beta_{2,i}^2\Delta t_i}{\sigma_u^2}}{\frac{1}{\sigma_s^2}+\sum_{i=1}^N\frac{\beta_{2,i}^2\Delta t_i}{\sigma_u^2}}s-\sum_{i=1}^N\frac{\rho\sigma_v}{\sigma_s}\frac{\frac{\beta_{2,i}^2\Delta t_i}{\sigma_u^2}}{\frac{1}{\sigma_s^2}+\sum_{i=1}^n\frac{\beta_{2,i}^2\Delta t_i}{\sigma_u^2}}s_{y,i}\\ &= \bar{v}+\frac{\rho\sigma_v}{\sigma_s}\left(s-\frac{\frac{1}{\sigma_s^2}s+\sum_{i=1}^N\frac{\beta_{2,i}^2\Delta t_i}{\sigma_u^2}s_{y,i}}{\frac{1}{\sigma_s^2}+\sum_{i=1}^N\frac{\beta_{2,i}^2\Delta t_i}{\sigma_u^2}}\right).\end{aligned}$$

Since the BLUE of $\bar{s}$ at t_N is

$$\hat{s}_N=\frac{\frac{1}{\sigma_s^2}s+\sum_{i=1}^N\frac{\beta_{2,i}^2\Delta t_i}{\sigma_u^2}s_{y,i}}{\frac{1}{\sigma_s^2}+\sum_{i=1}^N\frac{\beta_{2,i}^2\Delta t_i}{\sigma_u^2}},$$

we have $P_N^r(s,s_{y,1},\ldots,s_{y,N})=\mathbb{E}_{\hat{s}_N}[v|s,y_1,\ldots,y_N]$. One important thing to notice is that the coefficients $\eta_{N,1},\ldots,\eta_{N,N+2}$ in the optimal $P_N^r(s,s_{y,1},\ldots,s_{y,N})$ do not depend on any $\eta_{m,l}$ for any $m<N$ and $l\in\{1,\ldots,m+2\}$.

Consider the trading date t_{N-1}. The robustness problem is $\min\limits_{\eta_{N-1,1},\ldots,\eta_{N-1,N+1}}\max\limits_{\tilde{s}\in\mathbb{R}}$

$$\mathbb{E}_{\tilde{s}}\left[\begin{array}{c}c(|v-\eta_{N-1,1}-\eta_{N-1,2}s-\eta_{N-1,3}s_{y,1}-\cdots-\eta_{N-1,N+1}s_{y,N-1}|)\\ +c(|v-\eta_{N,1}-\eta_{N,2}s-\eta_{N,3}s_{y,1}-\cdots-\eta_{N,N+2}s_{y,N}|)\end{array}\right], \tag{C.7}$$

in which $\eta_{N,1},\ldots,\eta_{N,N+2}$ are the optimal ones we find above. Since they are inde-

pendent of $\eta_{N-1,1},\ldots,\eta_{N-1,N+1}$, the only way that $\eta_{N-1,1},\ldots,\eta_{N-1,N+1}$ may affect $c(|v-\eta_{N,1}-\eta_{N,2}s-\eta_{N,3}s_{y,1}-\cdots-\eta_{N,N+2}s_{y,N}|)$ is through $s_{y,N}$. However, recall that regardless of the past prices, $s_{y,N}$ is constructed in a way such that $s_{y,N}=\bar{s}-\frac{u_N}{\beta_{2,N}\Delta t_N}$, which is also independent of $\eta_{N-1,1},\ldots,\eta_{N-1,N+1}$. Therefore, solving (C.7) is the same as solving

$$\min_{\eta_{N-1,1},\ldots,\eta_{N-1,N+1}} \max_{\tilde{s}\in\mathbb{R}} \mathbb{E}_{\tilde{s}}[c(|v-\eta_{N-1,1}-\eta_{N-1,2}s-\eta_{N-1,3}s_{y,1}-\cdots-\eta_{N-1,N+1}s_{y,N-1}|)].$$

Then, following the same arguments, we know that $P^r_{N-1}(s,s_{y,1},\ldots,s_{y,N-1})$ $=\mathbb{E}_{\hat{s}_{N-1}}[v|s,y_1,\ldots,y_{N-1}]$, $\ldots$, $P^r_n(s,s_{y,1},\ldots,s_{y,n})=\mathbb{E}_{\hat{s}_n}[v|s,y_1,\ldots,y_n]$.

The proof above can be extended to the case in which $\beta_{2,m}=0$ for some $m\in\{1,\ldots,N\}$ in a manner similar to the case of $\alpha_3=0$ in the proof of Proposition IV.2. The BLUE of $\bar{s}$ will leave out every y_m and $s_{y,m}$ such that $\beta_{2,m}=0$. When $\beta_{2,m}=0$, including y_m or $s_{y,m}$ only adds the liquidity traders' order u_m into the price, which increases the expected loss. Last, if $\beta_{2,m}=0$ for all $m\in\{1,\ldots,N\}$, the BLUE of $\bar{s}$ will be s. □

Proof of Theorem IV.3. We proceed in several steps.

Step 1. We verify that on each trading date, the optimal GMM estimator is the unique BLUE of $\bar{s}$. At each t_n, we have

$$X_n(\bar{s},s,p_1,\ldots,p_{n-1})=\left(\alpha+\beta_{1,n}s-\beta_{2,n}\bar{s}+\sum_{i=1}^{n-1}\gamma_{i,n}p_i\right)\Delta t_n$$

due to the linearity assumption. If $\beta_{2,n}\neq 0$ for every n, similar to how we construct s_y from y in the proof of Lemma 1, we can construct the following variable for each n given any realization of the public signal and past prices:

$$s_{y,n}=\frac{X_n+u_n-(\alpha+\beta_{1,n}s+\sum_{i=1}^{n-1}\gamma_{i,n}p_i)\Delta t_n}{-\beta_{2,n}\Delta t_n}=\bar{s}-\frac{u_n}{\beta_{2,n}\Delta t_n}.$$

Again, we can define $s=\bar{s}+\varepsilon_s$ in which $\varepsilon_s\sim\mathcal{N}(0,\sigma_s^2)$, and $s_{y,n}=\bar{s}+\varepsilon_{s_{y,n}}$ in which

$\varepsilon_{s_{y,n}} = -\frac{u_n}{\beta_{2,n}\Delta t_n}$. Then, at each t_n, we have $(n+1)$ moment conditions

$$\mathbb{E}[s] = \bar{s},\ \mathbb{E}[s_{y,1} - \bar{s}] = 0,\ \ldots,\ \mathbb{E}[s_{y,n} - \bar{s}] = 0.$$

The optimal GMM weighting matrix is

$$\left(\mathbb{E}\left[\begin{bmatrix} \bar{s} - s \\ \bar{s} - s_{y,1} \\ \vdots \\ \bar{s} - s_{y,n} \end{bmatrix} \times \begin{bmatrix} \bar{s} - s \\ \bar{s} - s_{y,1} \\ \vdots \\ \bar{s} - s_{y,n} \end{bmatrix}' \right] \right)^{-1} = \begin{bmatrix} \frac{1}{\sigma_s^2} & 0 & \cdots & 0 \\ 0 & \frac{\beta_{2,1}^2 \Delta t_1}{\sigma_u^2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{\beta_{2,n}^2 \Delta t_n}{\sigma_u^2} \end{bmatrix},$$

which is diagonal. Therefore, we can rewrite the GMM minimization problem as $\min_{\hat{s}} \frac{1}{\sigma_s^2}(s - \hat{s})^2 + \sum_{i=1}^{n} \frac{\beta_{2,i}^2 \Delta t_i}{\sigma_u^2}(s_{y,i} - \hat{s})^2$, which can in turn be written as the following OLS problem:

$$\begin{bmatrix} \frac{1}{\sigma_s} s \\ \frac{\beta_{2,1}\sqrt{\Delta t_1}}{\sigma_u} s_{y,1} \\ \vdots \\ \frac{\beta_{2,n}\sqrt{\Delta t_n}}{\sigma_u} s_{y,n} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_s} \\ \frac{\beta_{2,1}\sqrt{\Delta t_1}}{\sigma_u} \\ \vdots \\ \frac{\beta_{2,n}\sqrt{\Delta t_n}}{\sigma_u} \end{bmatrix} \times \bar{s} + \begin{bmatrix} \frac{1}{\sigma_s} \varepsilon_s \\ \frac{\beta_{2,1}\sqrt{\Delta t_1}}{\sigma_u} \varepsilon_{s_{y,1}} \\ \vdots \\ \frac{\beta_{2,n}\sqrt{\Delta t_n}}{\sigma_u} \varepsilon_{s_{y,n}} \end{bmatrix}. \tag{C.8}$$

All of the error terms in (C.8) are independent and follow the standard normal distribution. According to the Gauss–Markov theorem, the OLS estimator of $\bar{s}$ from (C.8) is the unique BLUE of $\bar{s}$. Therefore,

$$\hat{s}_n = \frac{\frac{1}{\sigma_s^2} s + \sum_{i=1}^{n} \frac{\beta_{2,i}^2 \Delta t_i}{\sigma_u^2} s_{y,i}}{\frac{1}{\sigma_s^2} + \sum_{i=1}^{n} \frac{\beta_{2,i}^2 \Delta t_i}{\sigma_u^2}}. \tag{C.9}$$

This expression also covers the case in which $\beta_{2,i} = 0$ for some $i \leq n$. If $\beta_{2,i} = 0$, y_i should not appear in $\hat{s}_n$ because it contains no information about $\bar{s}$, and indeed "$\beta_{2,i}^2 s_{y,i}$"$= 0$ if $\beta_{2,i} = 0$.

Define $\omega_n = \frac{1}{\sigma_{\bar{s}}^2} + \sum_{i=1}^{n} \frac{\beta_{2,i}^2 \Delta t_i}{\sigma_u^2}$ and $\omega_0 = \frac{1}{\sigma_{\bar{s}}^2}$. We can write (C.9) recursively:

$$\hat{s}_n = \frac{\omega_{n-1}\hat{s}_{n-1} + \frac{\beta_{2,n}^2 \Delta t_n s_{y,n}}{\sigma_u^2}}{\omega_{n-1} + \frac{\beta_{2,n}^2 \Delta t_n}{\sigma_u^2}}.$$

Now, the best estimate of $\bar{s}$ at t_n only depends on the best estimate of $\bar{s}$ at t_{n-1} and the new total order y_n.

Step 2. We prove by induction that in any dynamic BLUE equilibrium, $\hat{s}_n, P_{n,\hat{s}_n}, X_n, \pi_n$ satisfy the first five equations of the theorem. We start from the last period. First, $P_{N,\hat{s}_N} = \mathbb{E}_{\hat{s}_N}[v|s, y_1, \ldots, y_n] = \mathbb{E}_{\hat{s}_N}[v|s]$. The first equality is from Definition IV.7. The second equality is because once the distribution is chosen, the orders will be interpreted by the market maker using the chosen distribution (as opposed to the unknown true distribution), and they do not provide any additional information about v on top of s. To see this, for a given distribution, orders are functions of s and hence are weakly less informative about v than s. Since

$$\mathbb{E}_{\hat{s}_N}[v|s] = \bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \hat{s}_N)$$

and $\hat{s}_N$ only depends on $\hat{s}_{N-1}$ and y_N, we know that the price at t_N can be written as $P_{N,\hat{s}_N} = \lambda_{N,1} + \lambda_{N,2}s + \lambda_{N,3}y_N + \lambda_{N,4}\hat{s}_{N-1}$. From the probabilistically informed trader's profit maximization problem, we know that $\lambda_{N,3} \neq 0$.

The profit at t_N is

$$\mathbb{E}_{\bar{s}}[\pi_N|s, \hat{s}_{N-1}] = \left[\bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}) - \lambda_{N,1} - \lambda_{N,2}s - \lambda_{N,3}X_N - \lambda_{N,4}\hat{s}_{n-1}\right] X_N,$$

and therefore the profit maximization problem yields

$$X_N = \frac{\bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}) - \lambda_{N,1} - \lambda_{N,2}s - \lambda_{N,4}\hat{s}_{n-1}}{2\lambda_{N,3}}.$$

Then, we can determine $s_{y,N}$ for the market maker:

$$s_{y,N} = s - \frac{2\lambda_{N,3}y_N - \bar{v} + \lambda_{N,1} + \lambda_{N,2}s + \lambda_{N,4}\hat{s}_{n-1}}{\frac{\rho\sigma_v}{\sigma_s}} = \bar{s} - \frac{2\lambda_{N,3}u_N}{\frac{\rho\sigma_v}{\sigma_s}}$$

and

$$\frac{\frac{\rho\sigma_v}{\sigma_s}}{2\lambda_{N,3}} = \beta_{2,N}\Delta t_N.$$

Therefore, the price at t_N is

$$\begin{aligned} P_{N,\hat{s}_N} &= \bar{v} + \frac{\rho\sigma_v}{\sigma_s}\left[s - \frac{\omega_{N-1}\hat{s}_{N-1} + \frac{\beta_N^2\Delta t_N}{\sigma_u^2}(s - \frac{2\lambda_{N,3}y_N - \bar{v} + \lambda_{N,1} + \lambda_{N,2}s + \lambda_{N,4}\hat{s}_{n-1}}{\frac{\rho\sigma_v}{\sigma_s}})}{\omega_{N-1} + \frac{\beta_N^2\Delta t_N}{\sigma_u^2}}\right] \\ &= \lambda_{N,1} + \lambda_{N,2}s + \lambda_{N,3}y_n + \lambda_{N,4}\hat{s}_{n-1}. \end{aligned}$$

By matching the coefficients of the equation above, we find that $\lambda_{N,1} = \bar{v}$, $\lambda_{N,2} = \frac{\rho\sigma_v}{\sigma_s}$, and $\lambda_{N,4} = -\frac{\rho\sigma_v}{\sigma_s}$. Therefore, we have

$$P_{N,\hat{s}_N} = \bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \hat{s}_{N-1}) + \lambda_{N,3}y_N = P_{N-1,\hat{s}_{N-1}} + \lambda_{N,3}y_N,$$

$$\hat{s}_N = \hat{s}_{N-1} - \frac{\sigma_s}{\rho\sigma_v}\lambda_{N,3}y_N,$$

$$X_N = \frac{\rho\sigma_v(\hat{s}_{N-1} - \bar{s})}{2\lambda_{N,3}} = \beta_{2,N}\Delta t_N(\hat{s}_{N-1} - \bar{s}),$$

and

$$\begin{aligned} \mathbb{E}_{\bar{s}}[\pi_N|s, \hat{s}_n] &= \left[\frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}) - \frac{\rho\sigma_v}{\sigma_s}(s - \hat{s}_{N-1}) - \frac{\frac{\rho\sigma_v}{\sigma_s}}{2\beta_{2,N}\Delta t_N}\beta_{2,N}\Delta t_N(\hat{s}_{N-1} - \bar{s})\right] \\ &\cdot\beta_{2,N}\Delta t_N(\hat{s}_{N-1} - \bar{s}) = \frac{\rho\sigma_v}{2\sigma_s}\beta_{2,N}\Delta t_N(\hat{s}_{N-1} - \bar{s})^2. \end{aligned}$$

Define $\lambda_N := \lambda_{N,3}$ and $\beta_N := \beta_{2,N}$. According to the definition of ω_N, we know that

$$\omega_N = \omega_{N-1} + \frac{\beta_N^2\Delta t_N}{\sigma_u^2}.$$

Therefore, all five equations are satisfied in the last period.

Next, assume we already know that at t_{n+1}

$$P_{n+1,\hat{s}_{n+1}} = P_{n,\hat{s}_n} + \lambda_{n+1} y_{n+1},$$

$$\hat{s}_{n+1} = \hat{s}_n - \frac{\sigma_s}{\rho\sigma_v}\lambda_{n+1} y_{n+1},$$

$$X_{n+1} = \beta_{n+1}\Delta t_{n+1}(\hat{s}_n - \bar{s}),$$

$$\mathbb{E}_{\bar{s}}[\pi_{n+1}|s, \hat{s}_n] = \alpha_n(\hat{s}_n - \bar{s})^2 + \delta_n,$$

and

$$\omega_n = \omega_{n-1} + \frac{\beta_n^2 \Delta t_n}{\sigma_u^2}.$$

At t_n, following a similar argument, we know that we can write the price as $P_{n,\hat{s}_n} = \lambda_{n,1} + \lambda_{n,2}s + \lambda_{n,3}y_n + \lambda_{n,4}\hat{s}_{n-1}$. Now, the probabilistically informed trader's expected profit $\mathbb{E}_{\bar{s}}[\pi_n|s, \hat{s}_{n-1}]$ is

$$\begin{aligned}
\mathbb{E}_{\bar{s}}[\pi_n|s, \hat{s}_{n-1}] &= \mathbb{E}_{\bar{s}}[\alpha_n(\hat{s}_n - \bar{s})^2 + \delta_n + (v - P_{n,\hat{s}_n})X_n|s, \hat{s}_{n-1}] \\
&= \left(\bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}) - \lambda_{n,1} - \lambda_{n,2}s - \lambda_{n,3}X_n - \lambda_{n,4}\hat{s}_{n-1}\right) X_n \\
&\quad + \alpha_n \left(\frac{\bar{v} - \lambda_{n,1} - \lambda_{n,2}s - \lambda_{n,3}X_n - \lambda_{n,4}\hat{s}_{n-1}}{\frac{\rho\sigma_v}{\sigma_s}} + s - \bar{s}\right)^2 + \delta_n.
\end{aligned}$$

The second equality relies on the observation that $\hat{s}_n = \frac{\bar{v} - P_{n,\hat{s}_n}}{\frac{\rho\sigma_v}{\sigma_s}} + s$. The solution to the profit maximization problem is

$$X_n = \frac{1 - \frac{2\alpha_n\lambda_{n,3}}{(\frac{\rho\sigma_v}{\sigma_s})^2}}{2\lambda_{n,3}\left(1 - \frac{\alpha_n\lambda_{n,3}}{(\frac{\rho\sigma_v}{\sigma_s})^2}\right)} \left[\bar{v} - \lambda_{n,1} - \lambda_{n,2}s - \lambda_{n,4}\hat{s}_{n-1} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s})\right],$$

in which we have used the second-order condition $\lambda_{n,3}\left(1 - \frac{\alpha_n\lambda_{n,3}}{(\frac{\rho\sigma_v}{\sigma_s})^2}\right) > 0$.

Using the probabilistically informed trader's optimal order, we know that

$$\begin{aligned} s_{y,n} &= -\frac{2\lambda_{n,3}\left(1-\frac{\alpha_n\lambda_{n,3}}{(\frac{\rho\sigma_v}{\sigma_s})^2}\right)}{\left[1-\frac{2\alpha_n\lambda_{n,3}}{(\frac{\rho\sigma_v}{\sigma_s})^2}\right]\frac{\rho\sigma_v}{\sigma_s}}y_n + \frac{\bar{v}-\lambda_{n,1}+(\frac{\rho\sigma_v}{\sigma_s}-\lambda_{n,2})s-\lambda_{n,4}\hat{s}_{n-1}}{\frac{\rho\sigma_v}{\sigma_s}} \\ &= \bar{s}-\frac{u_n}{\beta_{2,n}\Delta t_n}, \end{aligned}$$

$$\frac{2\lambda_{n,3}\left(1-\frac{\alpha_n\lambda_{n,3}}{(\frac{\rho\sigma_v}{\sigma_s})^2}\right)}{\left[1-\frac{2\alpha_n\lambda_{n,3}}{(\frac{\rho\sigma_v}{\sigma_s})^2}\right]\frac{\rho\sigma_v}{\sigma_s}} = \frac{1}{\beta_{2,n}\Delta t_n},$$

and

$$P_{n,\hat{s}_n} = \bar{v}+\frac{\rho\sigma_v}{\sigma_s}\left(s-\frac{\omega_{n-1}\hat{s}_{n-1}+(\omega_n-\omega_{n-1})s_{y_n}}{\omega_n}\right) = \lambda_{n,1}+\lambda_{n,2}s+\lambda_{n,3}y_n+\lambda_{n,4}\hat{s}_{n-1}.$$

Again, we match the coefficients and get $\lambda_{n,1}=\bar{v}$, $\lambda_{n,2}=\frac{\rho\sigma_v}{\sigma_s}$, and $\lambda_{n,4}=-\frac{\rho\sigma_v}{\sigma_s}$. Thus,

$$P_{n,\hat{s}_n} = \bar{v}+\frac{\rho\sigma_v}{\sigma_s}(s-\hat{s}_{n-1})+\lambda_{n,3}y_n = P_{n-1,\hat{s}_{n-1}}+\lambda_{n,3}y_n,$$

$$\hat{s}_n = \frac{\bar{v}-P_{n,\hat{s}_n}}{\frac{\rho\sigma_v}{\sigma_s}}+s = \frac{\bar{v}-P_{n-1,\hat{s}_{n-1}}-\lambda_{n,3}y_n}{\frac{\rho\sigma_v}{\sigma_s}}+s = \hat{s}_{n-1}-\frac{\sigma_s}{\rho\sigma_v}\lambda_{n,3}y_n,$$

$$X_n = \frac{1-\frac{2\alpha_n\lambda_{n,3}}{(\frac{\rho\sigma_v}{\sigma_s})^2}}{2\lambda_{n,3}\left(1-\frac{\alpha_n\lambda_{n,3}}{(\frac{\rho\sigma_v}{\sigma_s})^2}\right)}\frac{\rho\sigma_v}{\sigma_s}(\hat{s}_{n-1}-\bar{s}) = \beta_{2,n}\Delta t_n(\hat{s}_{n-1}-\bar{s}),$$

and

$$\begin{aligned} \mathbb{E}_{\bar{s}}[\pi_n|s,\hat{s}_{n-1}] &= \mathbb{E}_{\bar{s}}[(v-P_{n,\hat{s}_n})X_n+\alpha_n(\hat{s}_n-\bar{s})^2+\delta_n|\hat{s}_{n-1},s] \\ &= \left[\frac{\rho\sigma_v}{\sigma_s}(\hat{s}_{n-1}-\bar{s})-\lambda_{3,n}X_n\right]X_n+\alpha_n\left(\hat{s}_{n-1}-\bar{s}-\frac{\sigma_s}{\rho\sigma_v}\lambda_{n,3}X_n\right)^2 \\ &\quad +\alpha_n\frac{\sigma_s^2\lambda_{3,n}^2}{\rho^2\sigma_v^2}\sigma_u^2\Delta t_n+\delta_n. \end{aligned}$$

Define $\lambda_n := \lambda_{n,3}$ and $\beta_n := \beta_{2,n}$.

In the last equation about $\mathbb{E}_{\bar{s}}[\pi_n|s,\hat{s}_{n-1}]$, if we plug in $X_n = \beta_n \Delta t_n(\hat{s}_{n-1} - \bar{s})$, we can write the first two terms that involve X_n as $C_n(\hat{s}_{n-1} - \bar{s})^2$ for some constant C_n. Hence, we can write the profit as

$$\mathbb{E}_{\bar{s}}[\pi_n|s,\hat{s}_{n-1}] = \alpha_{n-1}(\hat{s}_{n-1} - \bar{s})^2 + \delta_{n-1}.$$

Similarly,

$$\omega_n = \omega_{n-1} + \frac{\beta_n^2 \Delta t_n}{\sigma_u^2}$$

follows from the definition of ω_n. Therefore, we have shown by induction that in any dynamic BLUE equilibrium, $\hat{s}_n, P_{n,\hat{s}_n}, X_n, \pi_n$ satisfy the first five equations of the theorem.

Step 3. We show that $\{\alpha_n, \beta_n, \lambda_n, \delta_n, \omega_n\}_{n=1}^N$ satisfy the difference equation system. We have

$$y_n = X_n + u_n = \beta_n(\hat{s}_{n-1} - \bar{s})\Delta t_n + u_n,$$
$$s_{y,n} = \bar{s} - \frac{u_n}{\beta_n \Delta t_n} = \hat{s}_{n-1} - \frac{y_n}{\beta_n \Delta t_n},$$

and

$$\begin{aligned}
P_{n,\hat{s}_n} &= \bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \hat{s}_n) = \bar{v} + \frac{\rho\sigma_v}{\sigma_s}\left(s - \frac{\omega_{n-1}\hat{s}_{n-1} + \frac{\beta_n^2 \Delta t_n s_{y,n}}{\sigma_u^2}}{\omega_{n-1} + \frac{\beta_n^2 \Delta t_n}{\sigma_u^2}}\right) \\
&= \bar{v} + \frac{\rho\sigma_v}{\sigma_s}\left(s - \frac{\omega_{n-1}\hat{s}_{n-1} + \frac{\beta_n^2 \Delta t_n (\hat{s}_{n-1} - \frac{y_n}{\beta_n \Delta t_n})}{\sigma_u^2}}{\omega_{n-1} + \frac{\beta_n^2 \Delta t_n}{\sigma_u^2}}\right) \\
&= \bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \hat{s}_{n-1}) + \frac{\rho\sigma_v}{\sigma_s}\frac{\frac{\beta_n}{\sigma_u^2}}{\omega_{n-1} + \frac{\beta_n^2 \Delta t_n}{\sigma_u^2}} y_n \\
&= P_{n-1\hat{s}_{n-1}} + \lambda_n y_n = \bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \hat{s}_{n-1}) + \lambda_n y_n.
\end{aligned}$$

Comparing the two equations above and using the equation $\omega_n = \omega_{n-1} + \frac{\beta_n^2 \Delta t_n}{\sigma_u^2}$, we

find that

$$\lambda_n = \frac{\rho\sigma_v}{\sigma_s} \frac{\frac{\beta_n}{\sigma_u^2}}{\omega_{n-1} + \frac{\beta_n^2 \Delta t_n}{\sigma_u^2}}$$

and

$$\lambda_n \beta_n \Delta t_n = \frac{\rho\sigma_v}{\sigma_s} \left(1 - \frac{\omega_{n-1}}{\omega_n}\right),$$

which is the first equation in the difference equation system.

Next, we analyze the optimal order X_n the probabilistically informed trader chooses at t_n. By maximizing the profit, we can get the second difference equation:

$$\begin{aligned}
\mathbb{E}_{\bar{s}}[\pi_n | s, \hat{s}_{n-1}] &= \mathbb{E}_{\bar{s}}[(v - P_{n,\hat{s}_n}) X_n + \alpha_n (\hat{s}_n - \bar{s})^2 + \delta_n | s, \hat{s}_{n-1}] \\
&= \left[\bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}) - P_{n-1,\hat{s}_{n-1}} - \lambda_n (X_n + u_n)\right] X_n \\
&\quad + \alpha_n \mathbb{E}_{\bar{s}}[(\hat{s}_n - \bar{s})^2 | s, \hat{s}_{n-1}] + \delta_n.
\end{aligned}$$

Plugging in equation $\hat{s}_n = \frac{\bar{v} - P_n}{\frac{\rho\sigma_v}{\sigma_s}} + s$, we can write the profit as a quadratic function of X_n:

$$\begin{aligned}
\mathbb{E}_{\bar{s}}[\pi_n | s, \hat{s}_{n-1}] &= \mathbb{E}_{\bar{s}}\left[\bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}) - P_{n-1,\hat{s}_{n-1}} - \lambda_n (X_n + u_n) | \hat{s}_{n-1}, s\right] X_n \\
&\quad + \alpha_n \mathbb{E}_{\bar{s}}\left[\left(\frac{\bar{v} - P_{n-1,\hat{s}_{n-1}} - \lambda_n (X_n + u_n)}{\frac{\rho\sigma_v}{\sigma_s}} + s - \bar{s}\right)^2 \middle| \hat{s}_{n-1}, s\right] + \delta_n.
\end{aligned}$$

The solution is

$$X_n = \frac{\bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}) - P_{n-1,\hat{s}_{n-1}} - 2\alpha_n \frac{\lambda_n}{\frac{\rho\sigma_v}{\sigma_s}} \left(\frac{\bar{v} - P_{n-1,\hat{s}_{n-1}}}{\frac{\rho\sigma_v}{\sigma_s}} + s - \bar{s}\right)}{2\lambda_n \left(1 - \frac{\alpha_n \lambda_n}{(\frac{\rho\sigma_v}{\sigma_s})^2}\right)},$$

along with the second-order condition $\lambda_n \left(1 - \frac{\alpha_n \lambda_n}{(\rho\sigma_v/\sigma_s)^2}\right) > 0$.

Recall that $P_{n-1,\hat{s}_{n-1}} = \bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \hat{s}_{n-1})$. We have

$$X_n = \frac{1 - \frac{2\alpha_n\lambda_n}{(\frac{\rho\sigma_v}{\sigma_s})^2}}{2\lambda_n\left(1 - \frac{\alpha_n\lambda_n}{(\frac{\rho\sigma_v}{\sigma_s})^2}\right)}\frac{\rho\sigma_v}{\sigma_s}(\hat{s}_{n-1} - \bar{s}) = \beta_n\Delta t_n(\hat{s}_{n-1} - \bar{s})$$

and

$$\beta_n\Delta t_n = \frac{\rho\sigma_v}{2\lambda_n\sigma_s}\frac{1 - \frac{2\alpha_n\lambda_n}{(\rho\sigma_v/\sigma_s)^2}}{1 - \frac{\alpha_n\lambda_n}{(\rho\sigma_v/\sigma_s)^2}}.$$

Back to the equation about the profit:

$$\begin{aligned}
\mathbb{E}_{\bar{s}}[\pi_n|s, \hat{s}_{n-1}] &= \mathbb{E}_{\bar{s}}[(v - P_{n,\hat{s}_n})X_n + \alpha_n(\hat{s}_n - \bar{s})^2 + \delta_n|\hat{s}_{n-1}, s] \\
&= \mathbb{E}_{\bar{s}}\left[\bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}) - P_{n-1,\hat{s}_{n-1}} - \lambda_n(X_n + u_n)|\hat{s}_{n-1}, s\right]X_n \\
&+ \alpha_n\mathbb{E}_{\bar{s}}\left[\left(\frac{\omega_{n-1}\hat{s}_{n-1} + \frac{\beta_n^2\Delta t_n}{\sigma_u^2}(\hat{s}_{n-1} - \frac{X_n+u_n}{\beta_n\Delta t_n})}{\omega_n} - \bar{s}\right)^2\middle|\hat{s}_{n-1}, s\right] + \delta_n \\
&= \left[(\frac{\rho\sigma_v}{\sigma_s} - \lambda_n\beta_n\Delta t_n)\beta_n\Delta t_n + \alpha_n\left(\frac{\omega_{n-1}}{\omega_{n-1} + \frac{\beta_n^2\Delta t_n}{\sigma u^2}}\right)^2\right] \\
&\quad \cdot(\hat{s}_{n-1} - \bar{s})^2 + \delta_n + \alpha_n\frac{\frac{\beta_n^2\Delta t_n}{\sigma_u^2}}{(\omega_{n-1} + \frac{\beta_n^2\Delta t_n}{\sigma u^2})^2}.
\end{aligned}$$

The last equation results from

$$X_n = \beta_n\Delta t_n(\hat{s}_{n-1} - \bar{s}) \text{ and } \beta_n\Delta t_n = \frac{\rho\sigma_v}{2\lambda_n\sigma_s}\frac{1 - \frac{2\alpha_n\lambda_n}{(\rho\sigma_v/\sigma_s)^2}}{1 - \frac{\alpha_n\lambda_n}{(\rho\sigma_v/\sigma_s)^2}}.$$

We also have

$$\mathbb{E}_{\bar{s}}[\pi_n|s, p_1, \ldots, p_{n-1}] = \alpha_{n-1}(\hat{s}_{n-1} - \bar{s})^2 + \delta_{n-1} \text{ and } \omega_n = \omega_{n-1} + \frac{\beta_n^2\Delta t_n}{\sigma_u^2}.$$

By matching the coefficients, we can find that

$$\alpha_{n-1} = \left(\frac{\omega_{n-1}}{\omega_n}\right)^2 \alpha_n + \frac{\rho\sigma_v}{\sigma_s}\frac{\omega_{n-1}}{\omega_n}\beta_n \Delta t_n$$

and

$$\delta_{n-1} = \delta_n + \frac{\alpha_n \beta_n^2 \Delta t_n}{\omega_n^2 \sigma_u^2}.$$

By definition, $\pi_{N+1} = 0$, which implies that the boundary condition should be $\alpha_N = \delta_N = 0$. We have defined $\omega_0 = \frac{1}{\sigma_s^2}$. Thus, we have shown that $\{\alpha_n, \beta_n, \lambda_n, \delta_n, \omega_n\}_{n=1}^N$ is a solution to the difference equation system subject to $\omega_0 = 1/\sigma_s^2$, $\alpha_N = \delta_N = 0$, and $\lambda_n \left(1 - \frac{\alpha_n \lambda_n}{(\rho\sigma_v/\sigma_s)^2}\right) > 0$.

Step 4. Last, we show that this difference equation system has a unique solution. Using the definition of ω_n and the first equation in the difference equation system, we have

$$\lambda_n \beta_n \Delta t_n = \frac{\rho\sigma_v}{\sigma_s}\left(1 - \frac{\omega_{n-1}}{\omega_n}\right) = \frac{\rho\sigma_v}{\sigma_s}\frac{\beta_n^2 \Delta t_n}{\omega_n \sigma_u^2}.$$

Using

$$\beta_n \Delta t_n = \frac{\lambda_n \omega_n \sigma_u^2 \sigma_s \Delta t_n}{\rho\sigma_v},$$

we have

$$\frac{\lambda_n \omega_n \sigma_u^2 \sigma_s \Delta t_n}{\rho\sigma_v} = \frac{\rho\sigma_v}{2\lambda_n \sigma_s}\frac{1 - \frac{2\alpha_n\lambda_n}{(\rho\sigma_v/\sigma_s)^2}}{1 - \frac{\alpha_n\lambda_n}{(\rho\sigma_v/\sigma_s)^2}}.$$

Rearranging it, we obtain

$$2\left(\frac{\sigma_s\sigma_u}{\rho\sigma_v}\right)^2 \omega_n \Delta t_n \frac{\alpha_n}{(\rho\sigma_v/\sigma_s)^2}\lambda_n^3 - 2\left(\frac{\sigma_s\sigma_u}{\rho\sigma_v}\right)^2 \omega_n \Delta t_n \lambda_n^2 - \frac{2\alpha_n\lambda_n}{(\rho\sigma_v/\sigma_s)^2} + 1 = 0. \quad \text{(C.10)}$$

We analyze the solution uniqueness and the signs of the parameters in the equation above by induction. First, we will focus on the last trading date and show that the solution is unique on the last trading date, and all the parameters are nonnegative. Then, we will show that if all the parameters at t_n are nonnegative and λ_n has a unique solution, the parameters at t_{n-1} are also nonnegative and λ_{n-1} also has a unique solution.

We start from t_N, in which $\alpha_N = 0$. Equation (C.10) reduces to

$$2\left(\frac{\sigma_s\sigma_u}{\rho\sigma_v}\right)^2 \omega_N \Delta t_N \lambda_N^2 = 1.$$

From the condition $\lambda_N\left(1 - \frac{\alpha_N\lambda_N}{(\rho\sigma_v/\sigma_s)^2}\right) > 0$, we know that $\lambda_N > 0$ when $\alpha_N = 0$. This means that equation (C.10) has a unique solution. Moreover, we know that $\beta_N\Delta t_N = \frac{\rho\sigma_v}{2\lambda_N\sigma_s} > 0$ and $\alpha_{N-1} = \left(\frac{\omega_{N-1}}{\omega_N}\right)^2 \alpha_N + \frac{\rho\sigma_v}{\sigma_s}\frac{\omega_{N-1}}{\omega_N}\beta_N\Delta t_N > 0$.

Next, suppose we have already shown that α_n is nonnegative. If $\alpha_n = 0$, the argument is the same as above. If $\alpha_n > 0$, define

$$f(\lambda_n) = 2\left(\frac{\sigma_s\sigma_u}{\rho\sigma_v}\right)^2 \omega_n\Delta t_n \frac{\alpha_n}{(\rho\sigma_v/\sigma_s)^2}\lambda_n^3 - 2\left(\frac{\sigma_s\sigma_u}{\rho\sigma_v}\right)^2 \omega_n\Delta t_n\lambda_n^2 - \frac{2\alpha_n\lambda_n}{(\rho\sigma_v/\sigma_s)^2} + 1.$$

From the condition $\lambda_n\left(1 - \frac{\alpha_n\lambda_n}{(\rho\sigma_v/\sigma_s)^2}\right) > 0$, we know that $0 < \lambda_n < \frac{(\rho\sigma_v/\sigma_s)^2}{\alpha_n}$. If we plug in $\lambda_n = 0$ and $\lambda_n = \frac{(\rho\sigma_v/\sigma_s)^2}{2\alpha_n}$, we will find that

$$f(0) = 1 > 0$$

and

$$f\left(\frac{(\rho\sigma_v/\sigma_s)^2}{2\alpha_n}\right) = -\left(\frac{\sigma_s\sigma_u}{\rho\sigma_v}\right)^2 \omega_n\Delta t_n\lambda_n^2 < 0.$$

As $\lambda_n \to +\infty$, this function diverges to $+\infty$; as $\lambda_n \to -\infty$, this function diverges to $-\infty$, which means that this function has a unique solution in the interval $\left(0, \frac{(\rho\sigma_v/\sigma_s)^2}{2\alpha_n}\right)$. Thus, this difference equation system has a unique solution. Moreover, we know that $\beta_n\Delta t_n = \frac{\rho\sigma_v}{2\lambda_n\sigma_s}\frac{1-\frac{2\alpha_n\lambda_n}{(\rho\sigma_v/\sigma_s)^2}}{1-\frac{\alpha_n\lambda_n}{(\rho\sigma_v/\sigma_s)^2}} > 0$ and $\alpha_{n-1} > 0$. Therefore, by induction, we have shown that the difference equation system has a unique solution and the signs of all parameters are nonnegative. □

Proof of Proposition IV.3. We prove this proposition by induction. First, suppose

$n = 1$. From our characterization of the dynamic BLUE equilibrium,

$$\begin{aligned}
\mathbb{E}_{\bar{s}}[P_{1,\hat{s}_1}(s,y)|s] &= \mathbb{E}_{\bar{s}}[P_{0,s}(s,y) + \lambda_1 y_1|s] \\
&= \bar{v} + \frac{\rho\sigma_v}{\sigma_s\beta_1\Delta t_1}\left(1 - \frac{\omega_0}{\omega_1}\right)\beta_1\Delta t_1(\hat{s}_0 - \bar{s}) \\
&= \bar{v} + \frac{\rho\sigma_v}{\sigma_s}\left(1 - \frac{\omega_0}{\omega_1}\right)(s - \bar{s}).
\end{aligned}$$

Next, suppose that at t_n, $\mathbb{E}_{\bar{s}}[P_{n,\hat{s}_n}(s,y)|s] = \bar{v} + \theta_n\frac{\rho\sigma_v}{\sigma_s}(s - \bar{s})$ with $\theta_n = 1 - \frac{\omega_0}{\omega_n}$. Then, at t_{n+1},

$$\begin{aligned}
\mathbb{E}_{\bar{s}}[P_{n+1,\hat{s}_{n+1}}(s,y)|s] &= \mathbb{E}_{\bar{s}}[P_{n,\hat{s}_n}(s,y) + \lambda_n y_n|s] \\
&= \bar{v} + \frac{\rho\sigma_v}{\sigma_s}\left(1 - \frac{\omega_0}{\omega_n}\right)(s - \bar{s}) \\
&\quad + \frac{\rho\sigma_v}{\sigma_s\beta_{n+1}\Delta t_{n+1}}\left(1 - \frac{\omega_n}{\omega_{n+1}}\right)\beta_{n+1}\Delta t_{n+1}(\hat{s}_n - \bar{s}) \\
&= \bar{v} + \frac{\rho\sigma_v}{\sigma_s}\left(1 - \frac{\omega_0}{\omega_n}\right)(s - \bar{s}) \\
&\quad + \frac{\rho\sigma_v}{\sigma_s}\left(1 - \frac{\omega_n}{\omega_{n+1}}\right)(\hat{s}_n - \bar{s}).
\end{aligned}$$

According to the equation $\mathbb{E}_{\bar{s}}[P_{n,\hat{s}_n}(s,y)|s] = \bar{v} + \left(1 - \frac{\omega_0}{\omega_n}\right)\frac{\rho\sigma_v}{\sigma_s}(s - \bar{s})$, we know that

$$s - \hat{s}_n = \left(1 - \frac{\omega_0}{\omega_n}\right)(s - \bar{s}) \Rightarrow \hat{s}_n - \bar{s} = \frac{\omega_0}{\omega_n}(s - \bar{s}),$$

$$\begin{aligned}
\mathbb{E}_{\bar{s}}[P_{n+1,\hat{s}_{n+1}}(s,y)|s] &= \bar{v} + \frac{\rho\sigma_v}{\sigma_s}\left(1 - \frac{\omega_0}{\omega_n} + \frac{\omega_{n+1} - \omega_n}{\omega_{n+1}}\frac{\omega_0}{\omega_n}\right)(s - \bar{s}) \\
&= \bar{v} + \frac{\rho\sigma_v}{\sigma_s}\left(1 - \frac{\omega_0}{\omega_{n+1}}\right)(s - \bar{s}),
\end{aligned}$$

and therefore

$$\theta_{n+1} = 1 - \frac{\omega_0}{\omega_{n+1}}.$$

□

Proof of Proposition IV.4. Combining the first two equations in the difference equation system with the (recursive) defintion of ω_n, we have

$$\lambda_n \beta_n \Delta t_n = \frac{\rho \sigma_v}{\sigma_s} \frac{\frac{\beta_n^2 \Delta t_n}{\sigma_u^2 \omega_{n-1}}}{\frac{\beta_n^2 \Delta t_n}{\sigma_u^2 \omega_{n-1}} + 1}$$

and

$$1 - \frac{2\alpha_n \lambda_n}{(\rho \sigma_v / \sigma_s)^2} = \frac{\beta_n^2 \Delta t_n}{\sigma_u^2 \omega_{n-1}}.$$

From the first three equations in the difference equation system and eliminating ω_n using its definition, we obtain

$$\alpha_{n-1} = \frac{(\rho \sigma_v / \sigma_s)^2}{4\lambda_n \left(1 - \frac{\alpha_n \lambda_n}{(\rho \sigma_v / \sigma_s)^2}\right)}$$

and

$$\frac{\alpha_n - \alpha_{n-1}}{\alpha_{n-1}} = \frac{4\alpha_n \lambda_n \left(1 - \frac{\alpha_n \lambda_n}{(\rho \sigma_v / \sigma_s)^2}\right)}{(\rho \sigma_v / \sigma_s)^2} - 1 = -\left(1 - \frac{2\alpha_n \lambda_n}{(\rho \sigma_v / \sigma_s)^2}\right)^2.$$

We can also rewrite the expression of $\beta_n \Delta t_n$ as follows:

$$\beta_n \Delta t_n = \frac{\rho \sigma_v}{2\lambda_n \sigma_s} \frac{1 - \frac{2\alpha_n \lambda_n}{(\rho \sigma_v / \sigma_s)^2}}{1 - \frac{\alpha_n \lambda_n}{(\rho \sigma_v / \sigma_s)^2}} = \frac{2\alpha_{n-1} \left(1 - \frac{2\alpha_n \lambda_n}{(\rho \sigma_v / \sigma_s)^2}\right)}{(\rho \sigma_v / \sigma_s)}.$$

Now, define $\Phi_n = \frac{4\alpha_n^2}{(\rho \sigma_v / \sigma_s)^2 \omega_n \sigma_u^2}$. We have

$$1 - \frac{2\alpha_n \lambda_n}{(\rho \sigma_v / \sigma_s)^2} = \frac{\beta_n^2 \Delta t_n}{\sigma_u^2 \omega_{n-1}} = \frac{\left(2\alpha_{n-1} \left(1 - \frac{2\alpha_n \lambda_n}{(\rho \sigma_v / \sigma_s)^2}\right)\right)^2}{(\rho \sigma_v / \sigma_s)^2 \Delta t_n \omega_{n-1} \sigma_u^2}$$

and

$$1 - \frac{2\alpha_n \lambda_n}{(\rho \sigma_v / \sigma_s)^2} = \frac{\Delta t_n}{\Phi_{n-1}}.$$

From the definition of ω_n, we know that

$$\omega_n = \omega_{n-1} + \frac{\beta_n^2 \Delta t_n}{\sigma_u^2},$$

$$\frac{\omega_{n-1}}{\omega_n} = \frac{1}{\frac{\beta_n^2 \Delta t_n}{\sigma_u^2 \omega_{n-1}} + 1} = \frac{1}{1 + \frac{\Delta t_n}{\Phi_{n-1}}},$$

and hence

$$\frac{\alpha_n}{\alpha_{n-1}} = 1 - \left(1 - \frac{2\alpha_n \lambda_n}{(\rho\sigma_v/\sigma_s)^2}\right)^2 = 1 - \left(\frac{\Delta t_n}{\Phi_{n-1}}\right)^2.$$

These equations give us

$$\begin{aligned}
\frac{\Phi_n}{\Phi_{n-1}} &= \frac{\alpha_n^2}{\alpha_{n-1}^2} \frac{\omega_{n-1}}{\omega_n} \\
&= \left[1 - \left(\frac{\Delta t_n}{\Phi_{n-1}}\right)^2\right]^2 \left(1 + \frac{\Delta t_n}{\Phi_{n-1}}\right)^{-1} \\
&= \left[1 - \left(\frac{\Delta t_n}{\Phi_{n-1}}\right)^2\right] \left(1 - \frac{\Delta t_n}{\Phi_{n-1}}\right)
\end{aligned}$$

and

$$\Phi_n - \Phi_{n-1} = -\Delta t_n - \frac{\Delta t_n^2}{\Phi_{n-1}} + \frac{\Delta t_n^3}{\Phi_{n-1}^2}.$$

The boundary condition is $\Phi_N = 0$. We iterate this difference equation for Φ_n backward, and a cubic equation must be solved at each step:

$$\Phi_{n-1}^3 - \Phi_n \Phi_{n-1}^2 - \Delta t_n \Phi_{n-1}^2 - \Delta t_n^2 \Phi_{n-1} + \Delta t_n^3 = 0.$$

Let $f(\Phi_{n-1}) = \Phi_{n-1}^3 - \Phi_n \Phi_{n-1}^2 - \Delta t_n \Phi_{n-1}^2 - \Delta t_n^2 \Phi_{n-1} + \Delta t_n^3$. We have

$$f(0) = \Delta t_n^3 > 0,$$

$$f(\Phi_n) = -\Delta t_n \Phi_n^2 - \Delta t_n^2 \Phi_n + \Delta t_n^3 < 0,$$

$$f(\Phi_n + \Delta t_n) = -\Phi_n \Delta t_n^2 < 0,$$

and

$$f(\Phi_n + \frac{5}{4}\Delta t_n) = \Delta t_n[\Delta t_n - \frac{1}{2}(\Phi_n + \frac{5}{4}\Delta t_n)]^2 > 0.$$

Because $\frac{\alpha_n}{\alpha_{n-1}} < 1$ and $\omega_n > \omega_{n-1}$, we have $\Phi_{n-1} > \Phi_n$. As $\Phi_{n-1} \to -\infty$, the function f diverges to $-\infty$.

Of the three roots of the cubic equation, only the one that lies between $\Phi_n + \Delta t_n$ and $\Phi_n + \frac{5}{4}\Delta t_n$ satisfies $\Phi_{n-1} > \Phi_n$. We have

$$-\frac{5}{4} < \frac{\Phi_n - \Phi_{n-1}}{\Delta t_n} < -1$$

and

$$\frac{\Phi_n - \Phi_{n-1}}{\Delta t_n} \to -1 \text{ as } \frac{\Phi_n}{\Delta t_n} \to \infty.$$

Then, we know that Φ_n's converge uniformly to a continuous-time version of Φ, $\Phi(t) = 1 - t$.

Next, we work on the limiting behavior of $\omega(t)$. First, we have

$$\frac{\omega_{n-1} - \omega_n}{\omega_n} = \frac{1}{1 + \frac{\Delta t_n}{\Phi_{n-1}}} - 1 = -\frac{\Delta t_n}{1 - t_n} + o(|\Delta t_n|).$$

The solution to this equation converges to the solution of the difference equation

$$\frac{\omega(t)'}{\omega(t)} = \frac{1}{1-t},$$

which is

$$\omega(t) = \frac{1}{\omega_0(1-t)}.$$

Therefore, we know that as $\Delta t_n \to 0$, on the last trading date ($t \to 1$), $\omega_N \to \infty$, and hence $\hat{s}_N \to \bar{s}$, which in turn implies that if the Δt_n's go to zero, $P_{N,\hat{s}_N}$ converges to $\mathbb{E}_{\bar{s}}[v|s]$. □

Proof of Proposition IV.5. Suppose that the pricing strategy is $P^b(s,y) = \lambda_1^b + \lambda_2^b s +$

$\lambda_3^b y$ and the trading strategy is $X(s,\bar{s}) = \alpha_1^b + \alpha_2^b s + \alpha_3^b \bar{s}$. Recall that

$$\begin{bmatrix} v \\ s \\ \bar{s} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \bar{v} \\ \mu_{\bar{s}} \\ \mu_{\bar{s}} \end{bmatrix}, \begin{bmatrix} \sigma_v^2 & \rho\sigma_v\sigma_s & 0 \\ \rho\sigma_v\sigma_s & \sigma_s^2 + \sigma_{\bar{s}}^2 & \sigma_{\bar{s}}^2 \\ 0 & \sigma_{\bar{s}}^2 & \sigma_{\bar{s}}^2 \end{bmatrix} \right).$$

First, we solve the profit maximization problem for the probabilistically informed trader. This step is the same as in Lemma 1, except that we use a different notation $\mathbb{E}[v|\bar{s},s]$ for the probabilistically informed trader's conditional expectation of v rather than $\mathbb{E}_{\bar{s}}[v|s]$. Of course, the conditional expectation has not changed:

$$\mathbb{E}[v|\bar{s},s] = \bar{v} + \frac{\rho\sigma_v}{\sigma_s}(s - \bar{s}).$$

Therefore, we know that $\alpha_1^b = \frac{\bar{v}-\lambda_1^b}{2\lambda_3^b}$, $\alpha_2^b = \frac{\rho\sigma_v/\sigma_s - \lambda_2^b}{2\lambda_3^b}$ and $\alpha_3^b = -\frac{\rho\sigma_v}{2\sigma_s\lambda_3^b}$.

Next, we calculate $\mathbb{E}[v|s,y]$. Notice that

$$\begin{bmatrix} v \\ s \\ y \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \alpha_2^b & \alpha_3^b \end{bmatrix} \begin{bmatrix} v \\ s \\ \bar{s} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \alpha_1^b + u \end{bmatrix}.$$

Thus, the joint distribution of $(v,s,y)'$ is multivariate normal:

$$\mathcal{N}\left(\begin{bmatrix} \bar{v} \\ \mu_{\bar{s}} \\ (\alpha_2^b + \alpha_3^b)\mu_{\bar{s}} + \alpha_1^b \end{bmatrix}, \begin{bmatrix} \sigma_v^2 & \rho\sigma_v\sigma_s & \alpha_2^b\rho\sigma_v\sigma_s \\ \rho\sigma_v\sigma_s & \sigma_s^2 + \sigma_{\bar{s}}^2 & \alpha_2^b\sigma_s^2 + (\alpha_2^b + \alpha_3^b)\sigma_{\bar{s}}^2 \\ \alpha_2^b\rho\sigma_v\sigma_s & \alpha_2^b\sigma_s^2 + (\alpha_2^b + \alpha_3^b)\sigma_{\bar{s}}^2 & (\alpha_2^b)^2\sigma_s^2 \\ & & +(\alpha_2^b + \alpha_3^b)^2\sigma_{\bar{s}}^2 + \sigma_u^2 \end{bmatrix} \right).$$

This implies that

$$\begin{aligned}
\mathbb{E}[v|s,y] &= \bar{v} + \begin{bmatrix} \rho\sigma_v\sigma_s \\ \alpha_2^b\rho\sigma_v\sigma_s \end{bmatrix}' \begin{bmatrix} \sigma_s^2+\sigma_{\bar{s}}^2 & \alpha_2^b\sigma_s^2+(\alpha_2^b+\alpha_3^b)\sigma_{\bar{s}}^2 \\ \alpha_2^b\sigma_s^2+(\alpha_2^b+\alpha_3^b)\sigma_{\bar{s}}^2 & (\alpha_2^b)^2\sigma_s^2+(\alpha_2^b+\alpha_3^b)^2\sigma_{\bar{s}}^2+\sigma_u^2 \end{bmatrix}^{-1} \\
&\quad \cdot \begin{bmatrix} s-\mu_{\bar{s}} \\ y-\mathbb{E}[y] \end{bmatrix} \\
&= \bar{v} + \frac{1}{\sigma_u^2(\sigma_s^2+\sigma_{\bar{s}}^2)+(\alpha_3^b)^2\sigma_s^2\sigma_{\bar{s}}^2} \begin{bmatrix} \rho\sigma_v\sigma_s \\ \alpha_2^b\rho\sigma_v\sigma_s \end{bmatrix}' \\
&\quad \cdot \begin{bmatrix} (\alpha_2^b)^2\sigma_s^2+(\alpha_2^b+\alpha_3^b)^2\sigma_{\bar{s}}^2+\sigma_u^2 & -\alpha_2^b\sigma_s^2-(\alpha_2^b+\alpha_3^b)\sigma_{\bar{s}}^2 \\ -\alpha_2^b\sigma_s^2-(\alpha_2^b+\alpha_3^b)\sigma_{\bar{s}}^2 & \sigma_s^2+\sigma_{\bar{s}}^2 \end{bmatrix} \begin{bmatrix} s-\mu_{\bar{s}} \\ y-\mathbb{E}[y] \end{bmatrix} \\
&= \bar{v} + \frac{1}{\sigma_u^2(\sigma_s^2+\sigma_{\bar{s}}^2)+(\alpha_3^b)^2\sigma_s^2\sigma_{\bar{s}}^2} \begin{bmatrix} \rho\sigma_v\sigma_s[\sigma_u^2+(\alpha_2^b+\alpha_3^b)\alpha_3^b\sigma_{\bar{s}}^2] \\ -\rho\sigma_v\sigma_s\alpha_3^b\sigma_{\bar{s}}^2 \end{bmatrix}' \\
&\quad \cdot \begin{bmatrix} s-\mu_{\bar{s}} \\ y-(\alpha_2^b+\alpha_3^b)\mu_{\bar{s}}-\alpha_1^b \end{bmatrix}.
\end{aligned}$$

We match this expression with $P^b(s,y) = \lambda_1^b + \lambda_2^b s + \lambda_3^b y$. We first match the coefficient of s and y, and obtain

$$-\frac{\rho\sigma_v\sigma_s\alpha_3^b\sigma_{\bar{s}}^2}{\sigma_u^2(\sigma_s^2+\sigma_{\bar{s}}^2)+(\alpha_3^b)^2\sigma_s^2\sigma_{\bar{s}}^2} = \lambda_3^b \qquad \text{and} \qquad \frac{\rho\sigma_v\sigma_s[\sigma_u^2+(\alpha_2^b+\alpha_3^b)\alpha_3^b\sigma_{\bar{s}}^2]}{\sigma_u^2(\sigma_s^2+\sigma_{\bar{s}}^2)+(\alpha_3^b)^2\sigma_s^2\sigma_{\bar{s}}^2} = \lambda_2^b.$$

By plugging in $\alpha_1^b = \frac{\bar{v}-\lambda_1^b}{2\lambda_3^b}$, $\alpha_2^b = \frac{\rho\sigma_v/\sigma_s-\lambda_2^b}{2\lambda_3^b}$, and $\alpha_3^b = -\frac{\rho\sigma_v}{2\sigma_s\lambda_3^b}$, we find that

$$\lambda_2^b = \frac{\rho\sigma_v\sigma_s}{\sigma_s^2+\sigma_{\bar{s}}^2} \qquad \text{and} \qquad \lambda_3^b = \frac{\rho\sigma_v}{2\sigma_u}\sqrt{\frac{\sigma_{\bar{s}}^2}{\sigma_s^2+\sigma_{\bar{s}}^2}}.$$

By matching the constant term, we get $\lambda_1^b = \bar{v} - \frac{\rho\sigma_v\sigma_s}{\sigma_s^2+\sigma_{\bar{s}}^2}\mu_{\bar{s}}$. Finally, plugging these

parameters into the pricing strategy, we have

$$\begin{aligned}\mathbb{E}[P^b(s,y)|\bar{s},s] &= \bar{v}+\left(\frac{1}{2}+\frac{\sigma_s^2}{2\left(\sigma_{\bar{s}}^2+\sigma_s^2\right)}\right)\frac{\rho\sigma_v}{\sigma_s}\left(s-\frac{\bar{s}\,/\sigma_s^2+\mu_{\bar{s}}\,/(\sigma_{\bar{s}}^2+\sigma_s^2)}{1\,/\sigma_s^2+1\,/(\sigma_{\bar{s}}^2+\sigma_s^2)}\right)\\ &= \bar{v}+\frac{1}{2}\frac{\rho\sigma_v}{\sigma_s}(s-\bar{s})+\frac{1}{2}\frac{\rho\sigma_v\sigma_s}{\sigma_{\bar{s}}^2+\sigma_s^2}(s-\mu_{\bar{s}}).\end{aligned}$$

□

Proof of Proposition IV.6. Suppose the market maker's estimator of $\bar{s}$ is $\hat{s}(s,y) = k_1 + k_2 s + k_3 s_y$, in which

$$s_y = \frac{\sigma_s}{\rho\sigma_v}\left(\bar{v}-\lambda_1+\left(\frac{\rho\sigma_v}{\sigma_s}-\lambda_2\right)s-2\lambda_3(X(\bar{s},s)+u)\right) = \bar{s}-\frac{2\lambda_3\sigma_s}{\rho\sigma_v}u.$$

Then, the market maker uses this $\hat{s}(s,y)$ to form the conditional expectation of v:

$$\begin{aligned}\mathbb{E}_{\hat{s}(s,y)}[v|s,y] &= \bar{v}+\frac{\rho\sigma_v}{\sigma_s}(s-\hat{s}(s,y))\\ &= \bar{v}+\frac{\rho\sigma_v}{\sigma_s}\left(s-k_1-k_2s-k_3s_y\right)\\ &= \bar{v}+\frac{\rho\sigma_v}{\sigma_s}\begin{bmatrix} s-k_1-k_2s \\ -k_3\frac{\sigma_s}{\rho\sigma_v}\left(\bar{v}-\lambda_1+\left(\frac{\rho\sigma_v}{\sigma_s}-\lambda_2\right)s-2\lambda_3(X(\bar{s},s)+u)\right)\end{bmatrix}\\ &= \bar{v}-\frac{\rho\sigma_v}{\sigma_s}k_1-k_3(\bar{v}-\lambda_1)+\frac{\rho\sigma_v}{\sigma_s}(1-k_2)s-k_3(\frac{\rho\sigma_v}{\sigma_s}-\lambda_2)s+2k_3\lambda_3 y\\ &= P_{\hat{s}(s,y)}(s,y)\\ &= \lambda_1+\lambda_2 s+\lambda_3 y.\end{aligned}$$

This implies that

$$\lambda_1 = \bar{v}-\frac{2\rho\sigma_v}{\sigma_s}k_1, \qquad \lambda_2 = \frac{\rho\sigma_v}{\sigma_s}(1-2k_2), \qquad k_3 = 1/2.$$

Therefore, the first statement of the proposition holds; that is, for any λ_1, λ_2, and λ_3, we can find k_1, k_2, and k_3 that satisfy the equilibrium conditions. We can compute

the conditionally expected profit of the probabilistically informed trader:

$$\begin{aligned}
\mathbb{E}_{\bar{s}}[\pi|s] &= \mathbb{E}[(v - P_{\hat{s}(s,y)}(s,y))x|s] \\
&= \mathbb{E}\left[\left(v - \bar{v} + \frac{2\rho\sigma_v k_1}{\sigma_s} - \frac{\rho\sigma_v}{\sigma_s}(1-2k_2)s - \lambda_3(x+u)\right)x \middle| s\right] \\
&= -\lambda_3 x^2 + \left[\frac{\rho\sigma_v}{\sigma_s}(s-\bar{s}) + \frac{2\rho\sigma_v}{\sigma_s}k_1 - \frac{\rho\sigma_v}{\sigma_s}(1-2k_2)s\right]x,
\end{aligned}$$

with the maximum conditionally expected profit being

$$\mathbb{E}_{\bar{s}}[\pi^\star|s] = \frac{\left[\frac{\rho\sigma_v}{\sigma_s}(s-\bar{s}) + \frac{2\rho\sigma_v}{\sigma_s}k_1 - \frac{\rho\sigma_v}{\sigma_s}(1-2k_2)s\right]^2}{4\lambda_3}.$$

Next, if we further require that $\hat{s}$ be unbiased, we will find that $k_1 = 0$, $k_2 = 1/2$, $\lambda_1 = \bar{v}$, and $\lambda_2 = 0$. Note that λ_3 can take any value. Thus, we have the second statement of the proposition. In addition, the optimal conditionally expected profit of the probabilistically informed trader becomes

$$\mathbb{E}_{\bar{s}}[\pi^\star|s] = \frac{\rho^2\sigma_v^2}{4\lambda_3\sigma_s^2}(s-\bar{s})^2.$$

Comparing this equation with the one under the BLUE, it is clear that whether the profit above is higher than the one under the BLUE depends on the value of λ_3. Since λ_3 can take any positive value, statements 3 and 4 follow. □

BIBLIOGRAPHY

BIBLIOGRAPHY

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies 72*(1), 1–19.

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017). When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research.

Abadie, A. and M. D. Cattaneo (2018). Econometric methods for program evaluation. *Annual Review of Economics 10*, 465–503.

Athey, S. and G. W. Imbens (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica 74*(2), 431–497.

Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments*, Volume 1, pp. 73–140. Elsevier.

Back, K., C. Cao, and G. Willard (2000). Imperfect Competition among Informed Traders. *Journal of Finance 55*(5), 2117–2155.

Ball, I. (2020). Scoring Strategic Agents. *Mimeo*.

Ball, R. and P. Brown (1968). An Empirical Evaluation of Accounting Income Numbers. *Journal of Accounting Research 6*(2), 159–178.

Beaver, W. (1968). The Information Content of Annual Earnings Announcements. *Journal of Accounting Research 6*(Empirical Research in Accounting: Selected Studies 1968), 67–92.

Benigno, P. and L. Paciello (2014). Monetary Policy, Doubts and Asset Prices. *Journal of Monetary Economics 64*, 85–98.

Bernard, V. (1992). *Advances in Behavioral Finance*, Chapter 11 Stock Price Reactions to Earnings Announcements: A Summary of Recent Anomalous Evidence and Possible Explanations, pp. 303–340. Russell Sage Foundation.

Bernard, V. and J. Thomas (1989). Post-Earnings-Announcement Drift: Delayed Price Response or Risk Premium? *Journal of Accounting Research 27*, 1–36.

Bernard, V. and J. Thomas (1990). Evidence That Stock Prices Do Not Fully Reflect the Implications of Current Earnings for Future Earnings. *Journal of Accounting and Economics 13*(4), 305–340.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics 119*(1), 249–275.

Blankespoor, E., E. deHaan, and I. Marinovic (2020). Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics 70*(2-3), 101344.

Borusyak, K. and X. Jaravel (2017). Revisiting event study designs. *Available at SSRN 2826228*.

Bose, S. and L. Renou (2014). Mechanism Design with Ambiguous Communication Devices. *Econometrica 82*(5), 1853–1872.

Bošković, B. and L. Nøstbakken (2018). How much does anticipation matter? evidence from anticipated regulation and land prices.

Caldentey, R. and E. Stacchetti (2010). Insider Trading With a Random Deadline. *Econometrica 78*(1), 245–283.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources 50*(2), 317–372.

Canay, I. A., A. Santos, and A. M. Shaikh (2021). The wild bootstrap with a "small" number of "large" clusters. *Review of Economics and Statistics 103*(2), 346–363.

Carroll, G. (2015). Robustness and Linear Contracts. *American Economic Review 105*(2), 536–563.

Carroll, G. (2017). Robustness and Separation in Multidimensional Screening. *Econometrica 85*(2), 453–488.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association 113*(523), 1350–1361.

Chan, L., N. Jegadeesh, and J. Lakonishok (1996). Momentum Strategies. *Journal of Finance 51*(5), 1681–1713.

Chen, Z. and L. Epstein (2002). Ambiguity, Risk, and Asset Returns in Continuous Time. *Econometrica 70*(4), 1403–1443.

Cheng, X. (2020). Relative Maximum Likelihood Updating of Ambiguous Beliefs. *Mimeo*.

Chu, Y., D. Hirshleifer, and L. Ma (2020). The Causal Effect of Limits to Arbitrage on Asset Pricing Anomalies. *Journal of Finance* (forthcoming).

Condie, S. and J. Ganguli (2011). Ambiguity and Rational Expectations Equilibria. *Review of Economic Studies 78*(3), 821–845.

Conley, T., S. Gonçalves, and C. Hansen (2018). Inference with dependent data in accounting and finance applications. *Journal of Accounting Research 56*(4), 1139–1203.

D'Adamo, R. (2019). Cluster-robust standard errors for linear regression models with many controls. *arXiv preprint arXiv:1806.07314*.

Daniel, K., D. Hirshleifer, and A. Subrahmanyam (1998). Investor Psychology and Security Market Under- and Overreactions. *Journal of Finance 53*(6), 1839–1885.

De Chaisemartin, C. and X. d'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–96.

Di Tillio, A., N. Kos, and M. Messner (2017). The Design of Ambiguous Mechanisms. *Review of Economic Studies 84*(1), 237–276.

Donoho, D. (1994). Statistical Estimation and Optimal Recovery. *Annals of Statistics 22*(1), 238–270.

Duffie, D. and P. Dworczak (2018). Robust Benchmark Design. *Mimeo*.

Easley, D. and M. O'Hara (2009). Ambiguity and Nonparticipation: The Role of Regulation. *Review of Financial Studies 22*(5), 1817–1843.

Easley, D. and M. O'Hara (2010). Microstructure and Ambiguity. *Journal of Finance 65*(5), 1817–1846.

Eliaz, K. and R. Spiegler (2018). The Model Selection Curse. *American Economic Review: Insights 1*(2), 127–140.

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics 75*(4), 643–669.

Epstein, L. and M. Schneider (2003). Recursive Multiple-Priors. *Journal of Economic Theory 113*(1), 1–31.

Epstein, L. and M. Schneider (2008). Ambiguity, Information Quality, and Asset Pricing. *Review of Financial Studies 63*(1), 197–228.

Epstein, L. and T. Wang (1994). Intertemporal Asset Pricing under Knightian Uncertainty. *Econometrica 62*(2), 283–322.

Esarey, J. and A. Menger (2019). Practical and effective approaches to dealing with clustered data. *Political Science Research and Methods 7*(3), 541–559.

Feingold, D. G. and R. S. Varga (1962). Block diagonally dominant matrices and generalizations of the gerschgorin circle theorem. *Pacific Journal of Mathematics 12*(4), 1241–1250.

Ferman, B. and C. Pinto (2019). Synthetic controls with imperfect pre-treatment fit. *arXiv preprint arXiv:1911.08521*.

Fitzpatrick, M. D. and M. F. Lovenheim (2014). Early retirement incentives and student achievement. *American Economic Journal: Economic Policy 6*(3), 120–54.

Foster, F. and S. Viswanathan (1996). Strategic Trading When Agents Forecast the Forecasts of Others. *Journal of Finance 51*(4), 1437–1478.

Frazzini, A. (2006). The Disposition Effect and Underreaction to News. *Journal of Finance 61*(4), 2017–2046.

Freyaldenhoven, S., C. Hansen, and J. M. Shapiro (2019). Pre-event trends in the panel event-study design. *American Economic Review 109*(9), 3307–38.

Galanis, S., C. Ioannou, and S. Kotronis (2019). Information Aggregation under Ambiguity: Theory and Experimental Evidence. *Mimeo*.

Gilboa, I. and D. Schmeidler (1989). Maxmin Expected Utility with Non-Unique Prior. *Journal of Mathematical Economics 18*(2), 141–153.

Gong, A. (2021). Bounds for treatment effects in the presence of anticipatory behavior. *arXiv preprint arXiv:2111.06573*.

Gong, A., S. Ke, Y. Qiu, and R. Shen (2022). Robust pricing under strategic trading. *Journal of Economic Theory 199*, 105201.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.

Greene, W. (2012). *Econometric Analysis* (8 ed.). Pearson.

Han, B., Y. Tang, and L. Yang (2016). Public Information and Uninformed Trading: Implications for Market Liquidity and Price Efficiency. *Journal of Economic Theory 163*(5), 604–643.

Hansen, B. E. and S. Lee (2019). Asymptotic theory for clustered samples. *Journal of Econometrics 210*(2), 268–290.

Hansen, L. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica 50*(4), 1029–1054.

Hansen, L. (2007). Beliefs, Doubts and Learning: Valuing Macroeconomic Risk. *American Economic Review 97*(2), 1–30.

Hansen, L. and T. Sargent (2001). Robust Control and Model Uncertainty. *American Economic Review 91*(2), 60–66.

Hansen, L. and T. Sargent (2008). *Robustness*. Princeton University Press.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: the LASSO and Generalizations*. CRC Press.

Heckman, J. J. and S. Navarro (2007). Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics 136*(2), 341–396.

Heckman, N. (1988). Minimax Estimates in a Semiparametric Model. *Journal of the American Statistical Association 83*(404), 1090–1096.

Hilary, G. and R. Shen (2013). The Role of Analysts in Intra-Industry Information Transfer. *The Accounting Review 88*(4), 1265–1287.

Hirano, K. and J. Porter (2003a). Asymptotic Efficiency in Parametric Structural Models with Parameter-Dependent Support. *Econometrica 71*(5), 1307–1338.

Hirano, K. and J. Porter (2003b). Efficiency in Asymptotic Shift Experiments. *Mimeo*.

Hirshleifer, D., S. Lim, and S. Teoh (2009). Driven to Distraction: Extraneous Events and Underreaction to Earnings News. *Journal of Finance 64*(5), 2289–2325.

Holden, C. and A. Subrahmanyam (1992). Long-Lived Private Information and Imperfect Competition. *Journal of Finance 47*(1), 247–270.

Holden, C. and A. Subrahmanyam (1994). Risk Aversion, Imperfect Competition, and Long-lived Information. *Economics Letters 44*(1–2), 181–190.

Hong, H., T. Lim, and J. Stein (2000). Bad News Travels Slowly: Size, Analyst Coverage, and the Profitability of Momentum Strategies. *Journal of Finance 55*(1), 265–295.

Hong, H. and J. Stein (1999). A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets. *Journal of Finance 54*(6), 2143–2184.

Hou, K., X. Chen, and L. Zhang (2018). Replicating Anomalies. *Review of Financial Studies* (forthcoming).

Hu, B. (2018). Statistical Arbitrage with Uncertain Fat Tails. *Mimeo*.

Huddart, S., J. Hughes, and C. Levine (2001). Public Disclosure and Dissimulation of Insider Trades. *Econometrica 69*(3), 665–681.

Ibragimov, R. and U. K. Müller (2016). Inference with few heterogeneous clusters. *Review of Economics and Statistics 98*(1), 83–96.

Ikenberry, D., J. Lakonishok, and T. Vermaelen (1995). Market Underreaction to Open Market Share Repurchases. *Journal of Financial Economics 39*(2–3), 181–208.

Imbens, G. W. and M. Kolesar (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics 98*(4), 701–712.

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*(6), 1845–1857.

Jéhiel, P. (2018). Investment Strategy and Selection Bias: An Equilibrium Perspective on Overoptimism. *American Economic Review 108*(6), 1582–1597.

Jochmans, K. (2020). Heteroscedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association*, 1–10.

Karantounias, A. (2013). Managing Pessimistic Expectations and Fiscal Policy. *Theoretical Economics 8*(1), 193–231.

Kilian, L. and H. Lütkepohl (2017). *Structural vector autoregressive analysis*. Cambridge University Press.

Klibanoff, P., M. Marinacci, and S. Mukerji (2005). A smooth Model of Decision Making under Ambiguity. *Econometrica 73*(6), 1849–1892.

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica, forthcoming*.

Kozlowski, J., L. Veldkamp, and V. Venkateswaran (2019). The Tail that Wags the Economy: Beliefs and Persistent Stagnation. *Journal of Political Economy* (forthcoming).

Kyle, A. (1985). Continuous Auctions and Insider Trading. *Econometrica 53*(6), 1315–1335.

Lambert, N., M. Ostrovsky, and M. Panov (2018). Strategic Trading in Informationally Complex Environments. *Econometrica 86*(4), 1119–1157.

Lehmann, E. and G. Casella (1998). *Theory of Point Estimation* (2 ed.). Springer-Verlag New York.

Levy, G. and R. Razin (2018). An Explanation-Based Approach to Combining Forecasts. *Mimeo*.

Li, K.-C. (1982). Minimaxity of the Method of Regularization of Stochastic Processes. *Annals of Statistics 10*(3), 937–942.

Liang, A. (2018). Games of Incomplete Information Played by Statisticians. *Mimeo*.

Liang, K.-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*(1), 13–22.

MacKinnon, J. G. (2019). How cluster-robust inference is changing applied econometrics. *Canadian Journal of Economics/Revue canadienne d'économique 52*(3), 851–881.

MacKinnon, J. G., M. Ø. Nielsen, and M. Webb (2022). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics, forthcoming*.

MacKinnon, J. G., M. Ø. Nielsen, M. Webb, et al. (2020). *Testing for the appropriate level of clustering in linear regression models*. Department of Economics, Queen's University.

Malani, A. and J. Reif (2015). Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform. *Journal of Public Economics 124*, 1–17.

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal 16*(1), S1–S23.

Manski, C. F. and J. V. Pepper (2013). Deterrence and the death penalty: Partial identification analysis using repeated cross sections. *Journal of Quantitative Criminology 29*(1), 123–141.

Manski, C. F. and J. V. Pepper (2018). How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics 100*(2), 232–244.

Miao, J. and A. Rivera (2016). Robust Contracts in Continuous Time. *Econometrica 84*(4), 1405–1440.

Michaely, R., R. Thaler, and K. Womack (1995). Price Reactions to Dividend Initiations and Omissions: Overreaction or Drift? *Journal of Finance 50*(2), 573–608.

Molinari, F. (2020). Microeconometrics with partial identification. *Handbook of econometrics 7*, 355–486.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences 10*, 1–51.

Ng, J., I. Tuna, and R. Verdi (2013). Management Forecast Credibility and Underreaction to News. *Review of Accounting Studies 18*(4), 956–986.

Orlik, A. and L. Veldkamp (2015). Understanding Uncertainty Shocks and the Role of Black Swans. *Mimeo*.

Roth, A. and J. Rambachan (2020). An honest approach to parallel trends. *Working Paper*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688.

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

Schmitt, B. A. (1992). Perturbation bounds for matrix square roots and pythagorean sums. *Linear Algebra and its Applications 174*, 215–227.

Shao, J. (2003). *Mathematical Statistics* (2 ed.). Springer-Verlag New York.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica 77*(4), 1299–1315.

Stoye, J. (2020). A simple, short, but never-empty confidence interval for partially identified parameters. *arXiv preprint arXiv:2010.10484*.

Subrahmanyam, A. (1991). Risk Aversion, Market Liquidity, and Price Efficiency. *Review of Financial Studies 4*(3), 417–441.

Sun, L. and S. Abraham (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*.

Varah, J. M. (1975). A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications 11*(1), 3–5.

Vazquez-Bare, G. (2021). Identification and estimation of spillover effects in randomized experiments. *Journal of Econometrics*.

Wald, A. (1950). *Statistical Decision Functions*. Wiley.

Wooldridge, J. (2016). *Introductory Econometrics: A Modern Approach* (6 ed.). Cengage Learning.

Yang, L. and H. Zhu (2019). Back-Running: Seeking and Hiding Fundamental Information in Order Flows. *Review of Financial Studies* (forthcoming).

Ye, T., L. Keele, R. Hasegawa, and D. S. Small (2021). A negative correlation strategy for bracketing in difference-in-differences. *arXiv preprint arXiv:2006.02423*.

Zinodiny, S., S. Rezaei, and S. Nadarajah (2017). Bayes Minimax Estimation of the Mean Matrix of Matrix-Variate Normal Distribution under Balanced Loss Function. *Statistics & Probability Letters 125*, 110–120.