Identification and estimation of dynamic random coefficient models*

Wooyong Lee†

May 2, 2022
(click here to view the latest version)

Abstract

This paper studies dynamic linear panel data models that allow multiplicative as well as additive heterogeneity in a short panel context, by allowing both the coefficients and intercept of linear models to be individual-specific. I show that the model is not point-identified and yet partially identified, and I characterize sharp identified sets of the mean, variance, and CDF of the partial effect distribution. The characterization applies to both discrete and continuous data. A computationally feasible estimation and inference procedure is proposed, based on a fast and exact global polynomial optimization algorithm. The method is applied to study lifecycle earnings dynamics in U.S. households in the Panel Study of Income Dynamics (PSID) dataset. Results suggest that there are large heterogeneity in earnings persistence and that the households experience weaker earnings persistence than what is reported in the literature on earnings dynamics.

*This is based on my PhD dissertation at the University of Chicago. I am indebted to Stéphane Bonhomme, Alexander Torgovitsky, and Guillaume Pouliot for guidance and support. I thank Manuel Arellano, Timothy Armstrong, Antonio Galvao, Greg Kaplan, Roger Koenker, Zhipeng Liao, Jack Light, Azeem Shaikh, Shuyang Sheng, Panagiotis Toulis, and Ying Zhu for helpful discussions and comments. I also thank seminar participants at the Econometrics Workshop and Econometrics Working Group at the University of Chicago.

†Economics Discipline Group, University of Technology Sydney. Email: wooyong.lee@uts.edu.au

1 Introduction

A common approach used with dynamic panel data linear models is to allow for fixed effects (Arellano and Bond, 1991; Blundell and Bond, 1998), which are individual-specific intercepts that allow for heterogeneity in levels of outcome among individuals of similar observable characteristics. A dynamic fixed effect model offers a flexible form of additive unobserved heterogeneity, which helps a researcher explore research questions, such as the effectiveness of a policy. The model is well-understood for short panel data (i.e., panel data with a small number of waves).

In addition to unobserved heterogeneity in levels, there is ample evidence that individuals have unobserved heterogeneity that interacts with observable characteristics. For example, firms have different levels of efficiency when using labor and capital, households have different levels of persistence in their earnings regarding their past earnings, and individuals have different levels of return to education. Such multiplicative heterogeneity is an essential mechanism for heterogeneous responses to exogenous shocks and policies, such as employment subsidies, income tax reform, and tuition subsidies. Multiplicative heterogeneity has a first-order influence on various economic models. For example, heterogeneity in earnings persistence governs heterogeneity in earnings risk that households experience, which is a fundamental motive for precautionary savings in the lifecycle model of consumption and savings behaviors.

This paper studies a dynamic linear panel data model that allows for both multiplicative and additive unobserved heterogeneity (i.e., a dynamic random coefficient model) in a short panel context. Consider a stylized example:

$$Y_{it} = \beta_{i0} + \beta_{i1} Y_{i,t-1} + \varepsilon_{it},$$

where all variables are scalars and ε_{it} is uncorrelated with the current history of Y_{it} (up to $t-1$) but correlated with its future values. In this model, both the coefficient (β_{i1}) and the intercept (β_{i0}) are individual-specific, reflecting multiplicative and additive unobserved heterogeneity. The model also allows lagged outcome $Y_{i,t-1}$ to be a regressor, reflecting dynamics. Analysis of this model is challenging in short panels since it is impossible to learn about individual values of the β_is with a small number of waves. This paper is first to propose general methods of identifying and estimating moments and distributions of such β_is.

Most research on random coefficient models with short panels focus on non-dynamic contexts (Chamberlain, 1992; Wooldridge, 2005; Arellano and Bonhomme, 2012; Graham and Powell, 2012), requiring that ε_{it} be uncorrelated with the entire history of regressors. This

implies that future values of regressors are uncorrelated with current outcomes, which is difficult to justify. For example, a firm's labor purchase decision next year might correlate with this year's output since the firm might learn about its own efficiency of labor from the output. A researcher might also be interested in the dynamics itself. For example, earnings persistence of a household is an important parameter since high earnings persistence makes earnings shocks last, which reduces a household's consumption smoothing ability and hence household welfare.

For random coefficient models with short panels in a dynamic context, a limited set of results is available. Chamberlain (1993) showed that the mean of β_is in dynamic random coefficient models is not point-identified, which implies that the mean of β_is is not estimable consistently. Arellano and Bonhomme (2012) showed that when the regressors are binary, the mean of β_is for some subpopulation is identifiable and hence consistently estimable, but they did not provide a general identification result that allows consistent estimation and inference.

This paper is first to present a general identification result for dynamic random coefficient models that allows consistent estimation and inference. Identification results for various features of β_is are presented, including the mean, variance, and CDF of β_is. This paper proposes a computationally feasible method of estimation and inference regarding the features of β_is, an essential step of which is to use a fast and exact algorithm for solving global polynomial optimization problems. The estimation method is then applied to learn about heterogeneity in lifecycle earnings dynamics across U.S. households in the Panel Study of Income Dynamics (PSID) dataset. The results of this paper are presented in three steps.

First, this paper shows that dynamic random coefficient models are partially identified, which implies finite bounds that can be placed on parameters of interest. Results are general in that they allow regressors and coefficients to be discrete or continuous. A key idea for the results is to recast the identification problem into a linear programming problem (Honoré and Tamer, 2006; Mogstad, Santos, and Torgovitsky, 2018; Torgovitsky, 2019), which becomes an infinite-dimensional problem when regressors or coefficients are continuous. I then use the dual representation of infinite-dimensional linear programming (Galichon and Henry, 2009; Schennach, 2014) to obtain sharp bounds for parameters of interest.

Second, I show that the sharp bounds can be computed fast and reliably by exploiting the linear structure of the model. Computing sharp bounds obtained from dual representation involves solving a nested optimization problem in which a researcher maximizes an objective function that contains another minimization problem. An important computational issue is that the inner minimization problem is a global minimization problem of a possibly non-convex function, for which standard global optimization procedures are infeasible because

the problem is nested and hence must be solved many times with precision. I show that for random coefficient models, the inner objective function is a polynomial. I then use a fast and exact algorithm to solve the global polynomial optimization problem, the semidefinite relaxation algorithm (Lasserre, 2010, 2015). Using this algorithm, sharp bounds for parameters of interest can be computed timely, and inferences about the bounds based on testing moment inequalities (Chernozhukov, Lee, and Rosen, 2013; Romano, Shaikh, and Wolf, 2014; Chernozhukov, Chetverikov, and Kato, 2019; Bai, Santos, and Shaikh, 2019) can also be performed in a computationally tractable way. For researchers interested in using the semidefinite relaxation approach to global polynomial optimization, I offer a general-purpose R package optpoly that implements the approach[1].

Third, I estimate a reduced-form lifecycle model of earnings dynamics. It serves as a key input in various economic models, including models of lifecycle earnings and consumption dynamics (Hall and Mishkin, 1982; Blundell, Pistaferri, and Preston, 2008; Blundell, Pistaferri, and Saporta-Eksten, 2016; Arellano, Blundell, and Bonhomme, 2017). The literature usually assumes no heterogeneity or observable heterogeneity in earnings. This paper investigates unobserved heterogeneity in household earnings, as in Guvenen (2007, 2009). He pointed out that considering unobserved heterogeneity in common trends of earnings[2] leads to earnings persistence estimate that is significantly smaller than what is reported on the literature on earnings dynamics, although his model did not allow for heterogeneity in earnings persistence itself. I estimate a model that allows for unobserved heterogeneity in both common trends and earnings persistence, finding large heterogeneity in earnings persistence with an average persistence being smaller than what is reported in the literature as pointed out by Guvenen (2007, 2009).

Results from this paper extend to moment equality models with unobservable quantities, which can be used to address a range of economic questions. For example, it can be applied to analysis of heterogeneous relationships between earnings and labor supply (Abowd and Card, 1989), or to production function estimation of firm-specific efficiency in labor and capital (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg, Caves, and Frazer, 2015).

The remainder of this paper is structured as follows. From Section 2 to Section 4, a dynamic random coefficient model is introduced and identification results from the model are presented. From Section 5 to Section 6, the estimation, inference and computation method for the parameters of interest is introduced, and their statistical properties are assessed. In Section 8, the method is applied to lifecycle earnings dynamics. Section 9 concludes.

[1] Available at `https://github.com/wooyong/optpoly`.

[2] It is called heterogeneous income profiles (HIP) in the literature.

2 Model and motivating examples

The dynamic random coefficient model is specified as:

$$Y_{it} = Z_{it}'\gamma_i + X_{it}'\beta_i + \varepsilon_{it}, \qquad t = 1, \ldots, T, \tag{1}$$

where i is an index of individuals, T is the length of panel data, $(Y_{it}, Z_{it}, X_{it}) \in \mathbb{R} \times \mathbb{R}^q \times \mathbb{R}^p$ are observed real vectors at time $t = 1, \ldots, T$, and $\varepsilon_{it} \in \mathbb{R}$ is an idiosyncratic error term at time t. Let $Y_i = (Y_{i1}, \ldots, Y_{iT})$ be the full history of $\{Y_{it}\}$ and $Y_i^t = (Y_{i1}, \ldots, Y_{it})$ be the history of $\{Y_{it}\}$ up to time t. Define X_i, X_i^t, Z_i, Z_i^t similarly. Assume:

$$\mathbb{E}(\varepsilon_{it} | \gamma_i, \beta_i, Z_i, X_i^t) = 0. \tag{2}$$

It assumes that the error term is mean independent of the full history of $\{Z_{it}\}$ but of current history of $\{X_{it}\}$; $\{Z_{it}\}$ is strictly exogenous and $\{X_{it}\}$ is sequentially exogenous. The presence of a sequentially exogenous regressor makes (1) a dynamic model.

The model is studied in a short panel context, which corresponds to the asymptotics that the number of individuals $N \to \infty$, but T is fixed. The random variables (γ_i, β_i), the random coefficients, have the same dimensions as (Z_{it}, X_{it}), and they can freely correlate among themselves and to observed data. (γ_i, β_i) are viewed as unobserved random variables that are i.i.d. across i with a common nonparametric distribution, which is the sense that a random coefficient model extends a fixed effects model.

Simplified notation is used throughout the paper. Let $W_i = (Y_i', Z_i', X_i')' \in \mathcal{W}$ be the vector of observables and $V_i = (\gamma_i', \beta_i')' \in \mathcal{V}$ be the vector of unobservables. Then, ε_{it} is understood as a deterministic function of (W_i, V_i) by the relationship $\varepsilon_{it} = Y_{it} - Z_{it}'\gamma_i - X_{it}'\beta_i$.

This paper considers parameter θ that has the form:

$$\theta = \mathbb{E}(m(Y_i, Z_i, X_i, \gamma_i, \beta_i)) = \mathbb{E}(m(W_i, V_i))$$

for some known function m. Theoretical results are presented for a generic Borel measurable m, but I focus on the case in which m is either a polynomial or an indicator function of V_i which has computationally feasible estimation and inference procedures. This choice of m includes many important parameters of interest. For example, θ can be an element of mean vector $\mathbb{E}(\beta_i)$ or an element of second moment $\mathbb{E}(\beta_i\beta_i')$. θ can also be the error variance $\mathbb{E}(\varepsilon_{it}^2)$ since

$$\varepsilon_{it}^2 = (Y_{it} - Z_{it}'\gamma_i - X_{it}'\beta_i)^2$$

is a quadratic polynomial in (γ_i, β_i). Alternatively, m can be indicator function $\mathbf{1}(\beta_i \leq b)$ for some b, in which case θ is

$$\theta = \mathbb{E}(\mathbf{1}(\beta_i \leq b)) = \mathbb{P}(\beta_i \leq b)$$

which is a CDF of β_i evaluated at b.

Example 1 (Household earnings)**.** One of the simplest examples of (1) is the AR(1) model with heterogeneous coefficient:

$$Y_{it} = \gamma_i + \beta_i Y_{i,t-1} + \varepsilon_{it}, \tag{3}$$

where all variables are scalars. This is a special case of (1), with $Z_{it} = 1$ and $X_{it} = Y_{i,t-1}$.

The AR(1) process is a popular choice for empirical specification of the lifecycle earnings process, with Y_{it} being the log-earnings net of demographic variables, which is an important input in the lifecycle model of consumption and savings behavior[3]. Specification of the earnings process has a first-order influence on the model outcome. Persistence of earnings (β_i) governs earnings risk experienced by households, which is a fundamental motive of precautionary savings. The literature usually models it as an AR(1) process with no coefficient heterogeneity or more simply as a unit root process, which is an AR(1) process with $\gamma_i = 0$ and $\beta_i = 1$. Guvenen (2007, 2009) pointed out that, if $\beta_i = \beta$ is assumed to be homogeneous and if the coefficients on demographic variables (age, in particular) are heterogeneous, β is estimated to be significantly less than 1. During application, I allow for heterogeneity in both β_i and coefficient on age. Studies that allow for coefficient heterogeneity include Browning, Ejrnaes, and Alvarez (2010) and Alan, Browning, and Ejrnæs (2018), with factor structure on the coefficients.

Example 2 (Household consumption behavior)**.** Consider a model of lifecycle consumption behavior:

$$C_{it} = \gamma_{i0} + \gamma_{i1} Y_{it} + \beta_i A_{it} + \nu_{it}, \tag{4}$$

where all variables are scalars, C_{it} is non-durable consumption, Y_{it} is earnings, and A_{it} is asset holdings at time t, all measured in logs and net of demographic variables. In the model, Y_{it} may be taken as strictly exogenous, meaning that the future earnings stream is unaffected by the current consumption choice. However, A_{it} must be taken as sequentially exogenous since past and future assets and consumptions interrelate through the intertemporal budget constraint.

(4) can be considered an approximation of the consumption rule derived from a structural model (Blundell, Pistaferri, and Saporta-Eksten, 2016). One parameter of interest in (4) is

[3] In the literature, it is standard to add a transitory shock to (3).

γ_{i1}, which represents the elasticity of consumption to earnings. This quantity measures a household's ability to smooth consumption against exogenous changes in earnings, such as exogenous earnings shocks, which is a determinant of a household's consumption smoothing ability and hence household welfare. Similar to the case of Example 1, the literature focuses on models with no coefficient heterogeneity[4].

Another parameter of interest is β_i, the elasticity of consumption to asset holdings, which measures a household's ability to smooth consumption against exogenous changes to assets. (4) allows a researcher to estimate the quantity while being agnostic about the evolution of assets over time (i.e., under non-parametric evolution of the assets).

Results from this paper also extend to a multivariate version of (1), the multivariate random coefficient model:

$$\mathbf{Y}_{it} = \mathbf{Z}_{it}'\gamma_i + \mathbf{X}_{it}'\beta_i + \mathbf{e}_{it},$$

where $\mathbf{Y}_{it}$ is a $D \times 1$ vector of response variables, $\mathbf{Z}_{it}$ is a $D \times q$ matrix of strictly exogenous regressors, $\mathbf{X}_{it}$ is a $D \times p$ matrix of sequentially exogenous regressors, and $\mathbf{e}_{it}$ is a $D \times 1$ vector of idiosyncratic error terms. Assume:

$$\mathbb{E}(\mathbf{e}_{it}|\gamma_i, \beta_i, \mathbf{Z}_i, \mathbf{X}_i^t) = 0,$$

which is a multivariate extension of (2).

Example 3 (Joint model of household earnings and consumption behavior)**.** A researcher can combine (3) and (4) in Examples 1 and 2 and consider a joint lifecycle model of earnings and consumption behavior. If I combine the time t consumption equation and the time $t+1$ earnings equation, I obtain multivariate random coefficient model:

$$\begin{aligned} C_{it} &= \gamma_{i1} + \gamma_{i2}Y_{it} + \beta_{i1}A_{it} + \nu_{it}, \\ Y_{i,t+1} &= \gamma_{i3} + \beta_{i2}Y_{it} + \varepsilon_{it}. \end{aligned}$$

This can be written in matrix form:

$$\begin{pmatrix} C_{it} \\ Y_{i,t+1} \end{pmatrix} = \begin{pmatrix} 1 & Y_{it} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \gamma_{i1} \\ \gamma_{i2} \\ \gamma_{i3} \end{pmatrix} + \begin{pmatrix} A_{it} & 0 \\ 0 & Y_{it} \end{pmatrix} \begin{pmatrix} \beta_{i1} \\ \beta_{i2} \end{pmatrix} + \begin{pmatrix} \nu_{it} \\ \varepsilon_{it} \end{pmatrix}.$$

[4]See Jappelli and Pistaferri (2010) for a survey.

3 Identification of means

This and following sections present identification results of the dynamic random coefficient model defined in (1). This section focuses on identification of the means of random coefficients, and the next presents a general identification result. Focusing on the mean allows presenting results using simple algebra.

Consider identifying a parameter that has the form:

$$\mu_e = \mathbb{E}(e_\gamma' \gamma_i + e_\beta' \beta_i) = \mathbb{E}(e' V_i)$$

where e_γ and e_β are real-valued vectors that the researcher chooses and $e \equiv (e_\gamma', e_\beta')'$. For example, if $e_\gamma = 0$ and $e_\beta = (1, 0, \ldots, 0)'$, then μ_e is the expectation of the first entry of β_i.

I present results regarding identification of μ_e in two subsections. The first subsection shows that μ_e is generally not point-identified. The following subsection shows that μ_e is partially identified.

3.1 Failure of point-identification

This subsection shows that μ_e is generally not point-identified, by considering a specific example of (1) and showing that μ_e is not point-identified in that example.

The example considered is the AR(1) model with heterogeneous coefficients in which two waves are observed:

$$Y_{it} = \gamma_i + \beta_i Y_{i,t-1} + \varepsilon_{it}, \qquad \mathbb{E}(\varepsilon_{it} | \gamma_i, \beta_i, Y_i^{t-1}) = 0, \tag{5}$$

for $t = 1, 2$, where all variables are scalar.

The following proposition states that $\mathbb{E}(\beta_i)$ is not point-identified in this model. The failure of point-identification implies that there is no consistent estimator for $\mathbb{E}(\beta_i)$.

Proposition 1. *Consider the model defined in (5). Assume that $(Y_{i0}, Y_{i1}, Y_{i2}, \gamma_i, \beta_i) \in \mathcal{C}$, where $\mathcal{C}$ is a compact subset of $\mathbb{R}^5$. Assume also that $(Y_{i0}, Y_{i1}, Y_{i2}, \gamma_i, \beta_i)$ are absolutely continuous with respect to the Lebesgue measure and that their joint density is strictly positive on $\mathcal{C}$, with a lower bound $b > 0$. Then, $\mathbb{E}(\beta_i)$ is not point-identified.*

Proof. See Appendix A.1. □

The same result holds when $(Y_{i0}, Y_{i1}, Y_{i2}, \gamma_i, \beta_i)$ is discrete and the number of support points of (γ_i, β_i) is not too small relative to that of (Y_{i0}, Y_{i1}, Y_{i2}). Chamberlain (1993) showed

that $\mathbb{E}(\beta_i)$ is not point-identified in (5) when the Y_{it}s are discrete and ε_{it} is mean independent of Y_i^{t-1}. Proposition 1 generalizes the result, showing that point-identification also fails with stronger assumptions and continuous data. The proof suggests that the result holds for any finite $T \geq 2$. Failure of point-identification in both discrete and continuous cases in the AR(1) model suggests that it is a general feature of dynamic random coefficient models.

The proof of Proposition 1 uses following lemma that is worth stated separately:

Lemma 1. *Under assumptions of Proposition 1, $\mathbb{E}(\beta_i)$ is point-identified if and only if there exists an unbiased estimator of β_i in individual time series, i.e. a function $S(Y_{i0}, Y_{i1}, Y_{i2})$ such that*

$$\mathbb{E}(S(Y_{i0}, Y_{i1}, Y_{i2})|\beta_i) = \beta_i$$

almost surely. When such S exists, $\mathbb{E}(\beta_i)$ is identified by $\mathbb{E}(\beta_i) = \mathbb{E}(S(Y_{i0}, Y_{i1}, Y_{i2}))$.

Proof. An immediate consequence of Proposition 7 in Appendix A.1. □

I then show that there is no such S, which proves Proposition 1. The intuition for Lemma 1 is that since the distribution of β_i is unrestricted, information on individual β_i can be obtained only from its individual time series. In long panel context, such information can be obtained by a time series estimator of β_i that is consistent as $T \to \infty$. In short panel context where T is fixed, information on β_i that is consistent as $T \to \infty$ is not reliable. Lemma 1 shows that unbiased information is the only reliable information on β_i for point-identification in short panels. Lemma 1 extends to a general case considered in the next section.

3.2 Partial identification

A natural question following Proposition 1 is whether the data are informative at all about $\mathbb{E}(\beta_i)$, or whether there is no information. This subsection shows that the data are informative about $\mathbb{E}(\beta_i)$. In particular, I show that μ_e is partially identified for any fixed e.

I show that there are finite bounds L and U such that:

$$L \leq \mu_e \leq U.$$

L and U are estimable with data, implying that lower and upper bounds of μ_e are consistently estimable. Dependence of L and U on e is suppressed in the notation.

Using the notation in Section 2 and letting $R_{it} = (Z_{it}', X_{it}')'$ be the vector of regressors at time t, I concisely write (1) and (2) as:

$$Y_{it} = R_{it}' V_i + \varepsilon_{it}, \quad t = 1, \ldots, T, \tag{6}$$

and

$$\mathbb{E}(\varepsilon_{it}|V_i, Z_i, X_i^t) = 0. \tag{7}$$

Recall that the parameter of interest, μ_e, is:

$$\mu_e = \mathbb{E}(e'V_i).$$

In this section and throughout the paper, I use unconditional moment restrictions that are implications of (7). It is known that the set of unconditional moment restrictions of the form

$$\mathbb{E}(g(V_i, Z_i, X_i^t)\varepsilon_{it}) = 0, \tag{8}$$

indexed by a suitable class of functions g, is equivalent to the conditional moment restriction in (7) (Bierens, 1990; Stinchcombe and White, 1998; Andrews and Shi, 2013). I choose the class to be the set of polynomial functions and use its finite subset for estimation and inference. Such finite subset of unconditional moment restrictions contains less information than (7), but it yields a computationally feasible estimation and inference procedure. Partial identification results based on (7) are established in Appendix B.

Consider the following assumptions:

Assumption 1. Random variables $(W_i, V_i)_{t=1}^T$ and $(\varepsilon_{it})_{t=1}^T$ satisfy (6).

Assumption 2. $\sum_{t=1}^T R_{it}R_{it}'$ is positive definite with probability 1.

Assumption 3. Random variables $(W_i, V_i)_{t=1}^T$ and $(\varepsilon_{it})_{t=1}^T$ satisfy, for all $t = 1, \ldots, T$,

$$\mathbb{E}((R_{it}'V_i)\varepsilon_{it}) = 0,$$
$$\mathbb{E}((Z_i', X_i^{t\prime})'\varepsilon_{it}) = 0.$$

Assumption 1 states that the dynamic random coefficient model is correctly specified. Assumption 2 is a no-multicollinearity assumption imposed on individual time series. This is stronger than the assumption that the expectation $\mathbb{E}(\sum_{t=1}^T R_{it}R_{it}')$ is positive definite, a standard assumption made when V_i is constant. A stronger assumption is required because the distribution of V_i is unrestricted and each V_i can only be learned from its individual data. If there are individuals whose data are not informative about V_i, then the data are not informative about $\mathbb{E}(e'V_i)$ because the missing V_i values might be arbitrarily large or small.

Assumption 3 states unconditional moment restrictions that are implications of (7). The first equation in Assumption 3 states that the "explained part" $(R_{it}'V_i)$ and the "error term"

(ε_{it}) are orthogonal. The second equation states that ε_{it} is orthogonal to the full history of Z_{it} and the current history of X_{it}.

The following theorem shows that μ_e is partially identified under Assumptions 1 to 3.

Theorem 1. *Suppose that Assumptions 1 to 3 hold. Let $\lambda_t \in \mathbb{R}$, and let μ_t be a real vector whose dimension is the same as $S_{it} = (Z_i', X_i^{t\prime})'$ for $t = 1, \ldots, T$. Let $\lambda \equiv (\lambda_1, \ldots, \lambda_T)$ and $\mu \equiv (\mu_1', \ldots, \mu_T')$. Then $L \leq \mu_e \leq U$ where*

$$L = \min_{\lambda<0,\ \mu} \mathbb{E}\left[\sum_{t=1}^{T} \mu_t' S_{it} Y_{it} + \frac{1}{4} B_i(\lambda, \mu)' \left(\sum_{t=1}^{T} \lambda_t R_{it} R_{it}'\right)^{-1} B_i(\lambda, \mu)\right]$$

and

$$U = \max_{\lambda>0,\ \mu} \mathbb{E}\left[\sum_{t=1}^{T} \mu_t' S_{it} Y_{it} + \frac{1}{4} B_i(\lambda, \mu)' \left(\sum_{t=1}^{T} \lambda_t R_{it} R_{it}'\right)^{-1} B_i(\lambda, \mu)\right]$$

where

$$B_i(\lambda, \mu) = e + \sum_{t=1}^{T} \lambda_t R_{it} Y_{it} - \sum_{t=1}^{T} R_{it} S_{it}' \mu_t.$$

These are the sharp bounds of μ_e under Assumptions 1 to 3.

Proof. See Appendix A.2. □

Since Assumption 3 is an implication of (7), L and U in Theorem 1 are non-sharp bounds of μ_e under Assumptions 1 and 2 and (7). Even though they are not sharp, they are given as solutions to optimization problems over the space of real vectors (λ, μ), from which I obtain a computationally feasible and scalable estimation and inference procedure. In contrast, the sharp bounds under (7) involve optimization over the space of functions, which is hard to deal with computationally.

L and U have closed-form expressions, but I do not display them here because (i) they are very complicated and (ii) they can be computationally more demanding than solving optimization problems since the expressions involve inversion of a big matrix. Instead, I present the following proposition, which gives simple closed-form expressions for a non-sharp bound.

Proposition 2. *Suppose Assumptions 1 to 3 hold, and let L and U be defined as in Theorem 1. For brevity of notation, define*

$$\mathcal{R}_i = \frac{1}{T} \sum_{t=1}^{T} R_{it} R_{it}' \quad \textit{and} \quad \mathcal{Y}_i = \frac{1}{T} \sum_{t=1}^{T} R_{it} Y_{it}.$$

Then $[L, U] \subseteq [\tilde{L}, \tilde{U}]$ *where*

$$[\tilde{L}, \tilde{U}] = \left[\tilde{V} - \frac{1}{2}\sqrt{\mathcal{E}\mathcal{D}},\ \ \tilde{V} + \frac{1}{2}\sqrt{\mathcal{E}\mathcal{D}}\right]$$

and

$$\begin{aligned}
\tilde{V} &= \frac{1}{2}\mathbb{E}(\mathcal{R}_i^{-1}\mathcal{Y}_i) + \frac{1}{2}\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i), \\
\mathcal{E} &= e'\mathbb{E}(\mathcal{R}_i^{-1})e - e'\mathbb{E}(\mathcal{R}_i)^{-1}e, \\
\mathcal{D} &= \mathbb{E}(\mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i) - \mathbb{E}(\mathcal{Y}_i)'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i),
\end{aligned}$$

where $\mathcal{E} \geq 0$ *and* $\mathcal{D} \geq 0$ *and they are zero if and only if* $\mathcal{R}_i$ *and* $\mathcal{R}_i^{-1}\mathcal{Y}_i$ *are degenerate across individuals, respectively.* $[\tilde{L}, \tilde{U}]$ *are sharp bounds of* μ_e *under Assumptions 1 and 2 and the following implication of Assumption 3:*

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}((R_{it}'V_i)\varepsilon_{it}) = 0, \\
\sum_{t=1}^{T} \mathbb{E}(R_{it}\varepsilon_{it}) = 0.
\end{aligned}$$

Proof. See Appendix A.3. □

The closed-form expressions in Proposition 2 give intuition for when L and U are finite. The expressions imply that L and U are finite as long as $\mathbb{E}(\mathcal{R}_i)$, $\mathbb{E}(\mathcal{Y}_i)$, $\mathbb{E}(\mathcal{R}_i^{-1}\mathcal{Y}_i)$ and $\mathbb{E}(\mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i)$ are finite. $\mathcal{R}_i$ is the design matrix for individual i and $\mathcal{R}_i^{-1}\mathcal{Y}_i$ is the OLS estimator of V_i from individual time series.

I now explain the intution behind Theorem 1, focusing on the upper bound U. For any (λ, μ), consider the quantity:

$$Q(\lambda, \mu, W_i, V_i) = e'V_i + \sum_{t=1}^{T} \lambda_t(R_{it}'V_i)\varepsilon_{it} + \sum_{t=1}^{T} \mu_t' S_{it}\varepsilon_{it}.$$

Dependence of Q on e is suppressed in the notation. It is possible to interpret Q as "Lagrangian"; it is a sum of $e'V_i$ and the moment functions in which $\{\lambda_t\}$ and $\{\mu_t\}$ are Lagrange multipliers. Note that $\mathbb{E}(Q) = \mathbb{E}(e'V_i)$ because the second and third terms have zero expectation by Assumption 3.

If I substitute $\varepsilon_{it} = Y_{it} - R_{it}V_i$ into Q and rearrange terms in V_i, I obtain expression:

$$Q(\lambda, \mu, W_i, V_i) = \sum_{t=1}^{T} \mu_t' S_{it} Y_{it} + \left[e + \sum_{t=1}^{T} \lambda_t R_{it} Y_{it} - \sum_{t=1}^{T} R_{it} S_{it}' \mu_t\right]' V_i - V_i' \left(\sum_{t=1}^{T} \lambda_t R_{it} R_{it}'\right) V_i.$$

This is a quadratic polynomial in V_i whose first and second derivatives are

$$\frac{dQ}{dV_i} = \left[e + \sum_{t=1}^{T} \lambda_t R_{it} Y_{it} - \sum_{t=1}^{T} R_{it} S_{it}' \mu_t \right] - 2 \left(\sum_{t=1}^{T} \lambda_t R_{it} R_{it}' \right) V_i$$

and

$$\frac{d^2 Q}{dV_i dV_i'} = -2 \left(\sum_{t=1}^{T} \lambda_t R_{it} R_{it}' \right).$$

If $\lambda_1, \ldots, \lambda_T > 0$, then the second derivative is a negative definite matrix, in which case Q attains a global maximum at the solution to the first-order condition $dQ/dV_i = 0$. Let $P = \max_{v \in \mathcal{V}} Q(\lambda, \mu, W_i, v)$ be the resulting maximum, which is only a function of (λ, μ, W_i) since V_i is "maximized out." The following identity holds:

$$P(\lambda, \mu, W_i) \geq Q(\lambda, \mu, W_i, V_i).$$

Considering expectation on both sides yields

$$\mathbb{E}(P(\lambda, \mu, W_i)) \geq \mathbb{E}(Q) = \mu_e$$

which shows that $\mathbb{E}(P)$ is an upper bound for μ_e for any (λ, μ) such that $\lambda > 0$. Since the inequality holds for any (λ, μ) such that $\lambda > 0$, it follows that

$$\min_{\lambda > 0,\ \mu} \mathbb{E}(P(\lambda, \mu, W_i)) \geq \mu_e.$$

The left-hand side coincides with U in Theorem 1, the sharp upper bound under Assumptions 1 to 3, which is estimable with data since it only involves W_i and not V_i. The sharp lower bound can be obtained by repeating the same argument with $\lambda < 0$.

4 Identification of higher order moments and the CDFs

This section presents a general partial identification result for dynamic random coefficient models. I consider a parameter of interest of the form

$$\theta = \mathbb{E}(m(W_i, V_i))$$

for some known function $m : \mathcal{W} \times \mathcal{V} \mapsto \mathbb{R}$. I assume unconditional moment restrictions:

Assumption 4. Random vectors (W_i, V_i) satisfy:

$$\mathbb{E}(\phi_k(W_i, V_i)) = 0, \quad k = 1, \ldots, K,$$

where the ϕ_ks are real-valued moment functions and K is the number of moment restrictions.

ε_{it} does not appear in Assumption 4 because ε_{it} is understood as a deterministic function of (W_i, V_i) by the relationship $\varepsilon_{it} = Y_{it} - R_{it}'V_i$.

Example 4. Consider identification of $\mathbb{E}(e'V_i)$ discussed in the previous section. Assumptions 1 and 3 imply $K = T + qT^2 + pT(T+1)/2$ moment conditions. The ϕ_ks for $k = 1, \ldots, T$ are

$$\phi_k(W_i, V_i) = (R_{ik}'V_i)(Y_{ik} - R_{ik}'V_i).$$

The ϕ_ks for $k > T$ are entries of the vectors

$$(Z_i', X_i^{t\prime})'(Y_i - R_{it}'V_i), \quad t = 1, \ldots, T$$

which is a $(qT + pt)$-dimensional vector for each t.

I characterize the identified set of θ under Assumptions 1 and 4. The approach is to recognize that the identified set can be characterized using linear programs. I then show that their dual programs yield a tractable characterization of the identified set.

Let $P_{W,V} \in \mathcal{M}_{W\times V}$ be a bounded and finitely additive signed Borel measure on $\mathcal{W} \times \mathcal{V}$ and $\mathcal{M}_{W\times V}$ be the linear space of such measures equipped with the total variation norm. Let P_W be the marginal distribution of W_i that the econometrician observes. Given the notation, the sharp identified set I of θ is *defined* by:

$$\begin{aligned} I \equiv \Bigg\{ \int m(w,v)dP \;\Bigg|\; & P \in \mathcal{M}_{W\times V}, \quad P \geq 0, \\ & \int dP = 1, \\ & \int \phi_k(w,v)dP = 0, \quad k = 1, \ldots, K, \\ & \int P(w, dv) = P_W(w) \text{ for all } w \in \mathcal{W} \Bigg\}. \end{aligned}$$

Dependence of I on m, P_W, ϕ_ks, and $\mathcal{M}_{W\times V}$ are suppressed in the notation.

I is the collection of all $\int m(W_i, V_i)dP$ values implied from P such that (i) P is a probability distribution of (W_i, V_i), (ii) P satisfies moment restrictions, and (iii) the marginal distribution of W_i implied from P equals the observed distribution P_W.

All defining properties of I are linear in P, which means that I is a convex set in $\mathbb{R}$ (i.e., an interval in $\mathbb{R}$). Therefore, I can be characterized by its lower and upper bounds. The sharp lower bound L of I is *defined* by:

$$\begin{aligned} \min_{P \in \mathcal{M}_{W \times V},\, P \geq 0} \int m(w,v) dP \qquad & \text{subject to} \\ \int \phi_k(w,v) dP = 0, \quad & k = 1, \ldots, K, \\ \int P(w, dv) = P_W(w) \;\; & \text{for all } w \in \mathcal{W}. \end{aligned} \tag{9}$$

The constraint $\int dP = 1$ is omitted since it is redundant given the last line of (9). $\int dP_W(w) = 1$ because it is a probability distribution.

Equation (9) is a linear program (LP) in P, with the caveat that P is an infinite-dimensional object. (9) is not a tractable characterization of L in the sense that the estimation methods that (9) imply are computationally infeasible. For example, Honoré and Tamer (2006) and Gunsilius (2019) discretized the space of (W_i, V_i) and solved the discretized problem, which is computationally infeasible for random coefficient models because the dimension of (W_i, V_i) is large. W_i contains the full history of regressors and response variables and V_i contains all random coefficients. For the random coefficient model with R regressors and T waves, P is a distribution on a $(RT + R + T)$-dimensional space.

My approach is to use the dual representation of (9) obtained by the duality theorem for infinite-dimensional LP (Galichon and Henry, 2009; Schennach, 2014). The duality theorem for infinite-dimensional LP extends that of a finite-dimensional LP. The following theorem characterizes I using the dual representation of (9) and the corresponding problem for the sharp upper bound.

Theorem 2. *Suppose Assumption 4 holds. Let $\lambda_k \in \mathbb{R}$ for $k = 1, \ldots, K$. Then, under suitable regularity conditions, including that $\mathcal{W} \times \mathcal{V}$ is compact and that $(m, \phi_1, \ldots, \phi_K)$ are bounded Borel measurable functions, $I = [L, U]$ where:*

$$L = \max_{\lambda_1, \ldots, \lambda_K} \mathbb{E}\left[\min_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} \right], \tag{10}$$

and

$$U = \min_{\lambda_1, \ldots, \lambda_K} \mathbb{E}\left[\max_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} \right]. \tag{11}$$

Proof. See Appendix A.4. □

Note that Theorem 2 does not include Assumption 1. Theorem 2 is a general duality result for models of moment equalities, where the moment functions contain both observables and unobservables (Galichon and Henry, 2009; Schennach, 2014; Chesher and Rosen, 2017; Li, 2018). In general, Theorem 2 leads to estimation and inference procedures that are not obvious to compute. However, I show in the next sections that, by exploiting linear structure of dynamic random coefficient models, I can obtain a computationally tractable estimation and inference procedure from this representation.

5 Estimation and inference

This section explains estimation and inference procedure for L and U defined in (10) and (11). This section focuses on describing the procedure, without discussing computation of objects involved in estimation and inference. The objects are not obvious to compute, and the next section discusses how to compute them.

5.1 Estimation

Theorem 2 characterizes the lower and upper bounds of θ in the population. In practice, a researcher does not observe the population distribution P_W but instead observes a finite sample $(W_1, \ldots, W_N)$ of size N which are i.i.d. P_W. A natural approach for estimating L and U given the sample is to replace expectations in (10) and (11) with sample means (the plug-in principle). Define $\hat{L}$ as an estimator for L where:

$$\hat{L} = \max_{\lambda_1,\ldots,\lambda_K} \frac{1}{N}\sum_{i=1}^{N} \min_{v\in\mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} \equiv \max_{\lambda\in\mathbb{R}^K} \frac{1}{N}\sum_{i=1}^{N} G_L(\lambda, W_i), \tag{12}$$

and $\hat{U}$ as an estimator for U where:

$$\hat{U} = \min_{\lambda_1,\ldots,\lambda_K} \frac{1}{N}\sum_{i=1}^{N} \max_{v\in\mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} \equiv \min_{\lambda\in\mathbb{R}^K} \frac{1}{N}\sum_{i=1}^{N} G_U(\lambda, W_i), \tag{13}$$

where $\lambda \in \mathbb{R}^K$. Let $[\hat{L}, \hat{U}]$ be the plug-in bound, given that they are formed using plug-in estimators.

The plug-in bound will be used as a key object for estimation and inference, but the bound is not obvious to compute. For example, consider computation of (12). The computation requires solving two types of optimization problems — solving the inner minimization problem

over $\mathcal{V}$ and solving the outer maximization problem with respect to $\lambda_1, \ldots, \lambda_T$. Each problem has its own difficulties:

- The inner minimization problem must be solved globally, but its objective function is not necessarily convex. It must also be solved *very fast*; it needs to be solved for each i and for each step of the outer maximization problem.
- The outer maximization problem must be solved globally, and it might be an optimization over a large dimensional space.

In addition, it will be shown that, for computational tractability of the outer problem, the inner problem must be solved not only very fast but also *exactly*. Thus, general-purpose global minimization methods that only provide approximate solution are not computationally feasible except for low-dimensional cases, such as $\mathcal{V}$ is discrete or $\mathcal{V} \subseteq \mathbb{R}^2$.

The next section discusses how to deal with these computational issues. In this section, I discuss estimation and inference given that we can compute the plug-in bound.

In what follows, I show consistency of lower plug-in bound in (12) to population lower bound in (10). The consistency of upper plug-in bound follows by the same argument, which yields consistency of the plug-in bound to population bound.

For lower plug-in bound, the solution function of the inner optimization problem

$$G_L(\lambda, w) = \min_{v \in \mathcal{V}} \left\{ m(w, v) + \sum_{k=1}^{K} \lambda_k \phi_k(w, v) \right\}$$

is a deterministic function given the model (i.e., given m and ϕ_ks) and the values (λ, w). Therefore, what is studied here is consistency of the statistical object

$$\hat{L} = \max_{\lambda} \hat{L}(\lambda) = \max_{\lambda} \frac{1}{N} \sum_{i=1}^{N} G_L(\lambda, W_i) \tag{14}$$

as an estimator for

$$L = \max_{\lambda} L(\lambda) = \max_{\lambda} \mathbb{E}\left(G_L(\lambda, W_i)\right). \tag{15}$$

$\hat{L}(\lambda)$ is the objective function of an M-estimation problem in which $L(\lambda)$ is the population objective and λ is the parameter that is M-estimated. Consistency then follows by replicating the analysis of M-estimation. The regularity conditions of M-estimation are satisfied by the fact that G_L is concave in λ, which will be shown in the next section.

Proposition 3. *Suppose that L exists and is finite, and that* $\operatorname{argmax}_\lambda L(\lambda)$ *is in the interior of* $\mathbb{R}^K$. *$\hat{L}$ then converges to L in probability.*

Proof. See Appendix A.5. □

5.2 Inference

This subsection discusses construction of a confidence interval for the identified set $[L, U]$ of $\theta \in \mathbb{R}$ given significance level α. The objective is to compute an interval $[L_\alpha, U_\alpha]$ such that:

$$\liminf_{N\to\infty} \inf_{P} \inf_{\theta\in[L,U]} P(\theta \in [L_\alpha, U_\alpha]) \geq 1 - \alpha.$$

Theorem 2 implies that any value $\theta \in [L, U]$ must satisfy

$$\theta \geq L = \max_{\lambda\in\mathbb{R}^K} \mathbb{E}(G_L(\lambda, W_i)),$$
$$\theta \leq U = \min_{\lambda\in\mathbb{R}^K} \mathbb{E}(G_U(\lambda, W_i)),$$

which implies

$$\theta \geq \mathbb{E}(G_L(\lambda, W_i)) \quad \text{for all } \lambda \in \mathbb{R}^K,$$
$$\theta \leq \mathbb{E}(G_U(\lambda, W_i)) \quad \text{for all } \lambda \in \mathbb{R}^K.$$

This implies following moment inequality conditions:

$$\begin{aligned} \mathbb{E}(G_L(\lambda, W_i) - \theta) \leq 0 \quad &\text{for all } \lambda \in \mathbb{R}^K, \\ \mathbb{E}(\theta - G_U(\lambda, W_i)) \leq 0 \quad &\text{for all } \lambda \in \mathbb{R}^K. \end{aligned} \tag{16}$$

(16) is a moment inequalities model with infinite number of moment restrictions (indexed by $\lambda \in \mathbb{R}^K$). I choose finite number of moment inequalities from (16) for computational tractability. Let Λ_F be a finite subset of $\mathbb{R}^K$ and consider a moment inequalities model with parameter θ_F:

$$\begin{aligned} \mathbb{E}(G_L(\lambda, W_i) - \theta_F) \leq 0 \quad &\text{for all } \lambda \in \Lambda_F, \\ \mathbb{E}(\theta_F - G_U(\lambda, W_i)) \leq 0 \quad &\text{for all } \lambda \in \Lambda_F. \end{aligned} \tag{17}$$

Since (17) uses a smaller number of moment inequalities than (16), a researcher can use (17) to make a conservative inference about θ in (16).

How conservative it is depends on how much information is contained in (17) relative to (16). Formal analysis of comparison between (16) and (17) is a topic of future research[5].

[5]Galichon and Henry (2011) studied reduction of the number of model restrictions without losing information. Their approach applies to the case in which the model outcomes, which are moment values in moment inequalities models, have discrete support.

There are two conjectures that can provide guidance. First, since G_L is concave in λ (and G_U is concave) and hence continuous in λ, I conjecture that setting Λ_F to be a grid of $\mathbb{R}^K$ yields a good approximation of (16). Second, concavity of G_L (and convexity of G_U) implies that two moments are binding in (16), namely the moments with indices $\lambda_L^* = \text{argmax}_\lambda \mathbb{E}(G_L(\lambda, W_i))$ and $\lambda_U^* = \text{argmin}_\lambda \mathbb{E}(G_U(\lambda, W_i))$. This means that it is sufficient to consider a grid around λ_L^* and λ_U^*[6]. The two conjectures lead to a strategy for how to choose Λ_F: estimate λ_L^* and λ_U^* using (12) and (13) and select points in the neighborhoods of them. I check performance of this strategy by simulation in later section.

Given Λ_F, (17) is a standard moment inequalities model except that Λ_F can be a large set. The literature on many moment inequalities (Romano, Shaikh, and Wolf, 2014; Bai, Santos, and Shaikh, 2019; Chernozhukov, Chetverikov, and Kato, 2019) propose how to compute a confidence interval $[L_\alpha, U_\alpha]$ that satisfies criterion for large Λ_F:

$$\liminf_{N\to\infty} \inf_P \inf_{\theta\in[L_F,U_F]} P(\theta \in [L_\alpha, U_\alpha]) \geq 1-\alpha.$$

Among the proposed methods, a procedure based on multiplier bootstrap by Chernozhukov, Chetverikov, and Kato (2019) is particularly appealing because of low computational cost of bootstrap. Their procedure uses following test statistic, computed for each $\theta_F \in \mathbb{R}$:

$$T_{CCK}(\theta_F) = \max\left\{\max_{\lambda\in\Lambda_F}\left\{\frac{\sqrt{N}(\mu_{G_L}(\lambda)-\theta_F)}{\sigma_{G_L}(\lambda)}\right\}, \max_{\lambda\in\Lambda_F}\left\{\frac{\sqrt{N}(\theta_F-\mu_{G_U}(\lambda))}{\sigma_{G_U}(\lambda)}\right\}\right\}$$

where

$$\mu_{G_L}(\lambda) = \frac{1}{N}\sum_{i=1}^N G_L(\lambda, W_i) \quad \text{and} \quad \sigma^2_{G_L}(\lambda) = \frac{1}{N}\sum_{i=1}^N \left(G_L(\lambda, W_i) - \mu_{G_L}(\lambda)\right)^2,$$

and $\mu_{G_U}(\lambda)$ and $\sigma^2_{G_U}(\lambda)$ are defined similarly with G_U.

T_{CCK} is then compared to a critical value $c_{CCK}(\alpha)$, computed using multiplier bootstrap. Each multiplier bootstrap replication simulates independent standard normal random draws $e_1, \ldots, e_N \in \mathbb{R}$ and computes the statistic:

$$c_{CCK} = \max\left\{\max_{\lambda\in\Lambda_F}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^N e_i\frac{G_L(\lambda, W_i)-\mu_{G_L}(\lambda)}{\sigma_{G_L}(\lambda)}\right\}, \max_{\lambda\in\Lambda_F}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^N e_i\frac{\mu_{G_U}(\lambda)-G_U(\lambda, W_i)}{\sigma_{G_U}(\lambda)}\right\}\right\}.$$

The critical value $c_{CCK}(\alpha)$ is then the $100\times(1-\alpha)$ percentile of the bootstrapped c_{CCK} values.

[6]This relates to a step in the inference procedure of Chernozhukov, Lee, and Rosen (2013), in which they compute a set of moment restrictions that are likely to bind.

The confidence interval is the set of θ_F such that

$$T_{CCK}(\theta_F) \leq c_{CCK}(\alpha). \tag{18}$$

In general, the procedure requires grid search over θ_F evaluating (18). In our case, two observations facilitate efficient search of θ_F. First, $c_{CCK}(\alpha)$ does not depend on θ_F, because c_{CCK} does not. This means $c_{CCK}(\alpha)$ is computed only once and is fixed when evaluating (18) for different θ_F. Second, if $\sigma_{G_L}(\lambda), \sigma_{G_U}(\lambda) > 0$ for all $\lambda \in \Lambda_F$ and confidence interval is not empty, then $T_{CCK}(\theta_F)$ is continuous in θ_F and the confidence interval is given by $[L_\alpha, U_\alpha]$ such that

$$T_{CCK}(L_\alpha) = c_{CCK}(\alpha) = T_{CCK}(U_\alpha)$$

where $L_\alpha \leq U_\alpha$. This leads to an efficient algorithm for computing a confidence interval. First, perform a crude grid search to find any interior point of the confidence interval, i.e. a point θ_F^* such that $T_{CCK}(\theta_F^*) < c_{CCK}(\alpha)$. Then solve the equation $T_{CCK}(\theta) = c_{CCK}(\alpha)$ on $(-\infty, \theta_F^*)$ to find L_α and on (θ_F^*, ∞) to find U_α, by minimizing squared distance $(T_{CCK}(\theta) - c_{CCK}(\alpha))^2$.

The above inference procedure naturally extends to a vector of parameters $\theta \in \mathbb{R}^d$, by considering (16) for every entry of θ. For example, the moment inequalities model for $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ is:

$$\begin{aligned}
\mathbb{E}(G_{L1}(\lambda, W_i) - \theta_1) &\leq 0 \quad \text{for all } \lambda \in \mathbb{R}^K, \\
\mathbb{E}(\theta_1 - G_{U1}(\lambda, W_i)) &\leq 0 \quad \text{for all } \lambda \in \mathbb{R}^K, \\
\mathbb{E}(G_{L2}(\lambda, W_i) - \theta_2) &\leq 0 \quad \text{for all } \lambda \in \mathbb{R}^K, \\
\mathbb{E}(\theta_2 - G_{U2}(\lambda, W_i)) &\leq 0 \quad \text{for all } \lambda \in \mathbb{R}^K,
\end{aligned} \tag{19}$$

where G_{Uk} and G_{Lk} are G_L and G_U in (16) for θ_k, $k = 1, 2$. Inference can then be performed by the same procedure, giving a confidence region in $\mathbb{R}^2$.

The above inference procedure calculates $G_L(\lambda, W_i)$ and $G_U(\lambda, W_i)$ for each $i \in N$ and $\lambda \in \Lambda_F$ only once for critical value calculation, a benefit of using multiplier bootstrap. A standard empirical bootstrap is computationally more expensive and usually infeasible. Each empirical bootstrap replication samples the data $(W_1^*, \ldots, W_N^*)$ with replacement and then computes:

$$\max\left\{ \max_{\lambda \in \Lambda_F} \left\{ \frac{\frac{1}{\sqrt{N}} \sum_{i=1}^N [G_L(\lambda, W_i^*) - \mu_{G_L}(\lambda)]}{\sigma_{G_L}(\lambda)} \right\}, \max_{\lambda \in \Lambda_F} \left\{ \frac{\frac{1}{\sqrt{N}} \sum_{i=1}^N [\mu_{G_U}(\lambda) - G_U(\lambda, W_i^*)]}{\sigma_{G_U}(\lambda)} \right\} \right\}.$$

The critical value is then the $100 \times (1 - \alpha)$ percentile of the bootstrapped values. This re-

quires calculating $G_L(\lambda, W_i^*)$ and $G_U(\lambda, W_i^*)$ for each $i \in N$ and $\lambda \in \Lambda_F$ for each bootstrap replication, which is computationally infeasible as the number of replications is large.

5.3 Estimation under empty plug-in bound

Although $\hat{L}$ is a consistent estimator for L, $\hat{L}$ is not always well-defined in the sample. The reason why can be understood by comparing it to a generalized method of moments (GMM) estimation problem. In GMM estimation, the minimum GMM objective might be strictly positive in the sample because the moment conditions might not be exactly satisfied in the sample. This also occurs with dynamic random coefficient models: there might be no distribution of the random coefficients that satisfies all moment conditions in the sample. In this case, the researcher obtains an empty identified set as a plug-in estimate, in which case the maximization problem of $\hat{L}$ diverges to $+\infty$ and the corresponding problem for the upper bound diverges to $-\infty$.

There are two approaches for resolving this issue. First, a researcher may not insist obtaining a point estimate, directly performing inference about the parameter without estimation. Second, a researcher may obtain a point estimate that minimizes distance between the model and the data. I first introduce the second approach in this subsection. The next subsection discusses the first approach which uses the second approach as a building block that has no statistical interpretation. For dynamic random coefficient models, I propose using the first approach. I consider this subsection as a building block for inference procedure in the next subsection.

In GMM estimation, when the moment conditions are not exactly satisfied, a researcher defines an estimator as the parameter value that minimizes the GMM criterion that is positive. A similar approach can be implemented for dynamic random coefficient models in two steps[7]. In the first step, the researcher finds the smallest $\delta \geq 0$ that satisfies criterion:

$$|\mathbb{E}(\phi_k(W_i, V_i))| \leq \delta, \quad k = 1, \ldots, K. \tag{20}$$

This can be thought of as an absolute-value GMM criterion. The following proposition explains how to compute the smallest δ.

[7] Andrews and Kwon (2019) study and formalize this approach for standard GMM estimation based on moments without unobservables.

Proposition 4. *Given the sample* $(W_1, \ldots, W_N)$*, consider linear programming problem:*

$$\min_{P \in \mathcal{M}_{W \times V},\ P \geq 0,\ \delta \geq 0} \delta \qquad \text{subject to} \qquad \begin{aligned} &\left| \int \phi_k(W_i, V_i) dP \right| \leq \delta, \quad k = 1, \ldots, K, \\ &\int P(w, dV_i) = \hat{P}_W(w) \ \text{ for all } w \in \mathcal{W}, \end{aligned} \tag{21}$$

where $\hat{P}_W$ *is the empirical distribution of* W_i *constructed from* $(W_1, \ldots, W_N)$*. Its solution then equals the solution to optimization problem:*

$$\max_{\lambda_1, \ldots, \lambda_K} \frac{1}{N} \sum_{i=1}^{N} \min_{v \in \mathcal{V}} \left\{ \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} \qquad \text{subject to} \qquad \sum_{k=1}^{K} |\lambda_k| \leq 1. \tag{22}$$

Proof. See Appendix A.6. □

Proposition 4 shows that a researcher can find the minimum δ by solving (22), which is similar to (12). One difference is that (22) is a constrained optimization problem, but the constraint has a very simple structure and its Jacobian can also be derived in closed-form.

Let δ^* be the solution to (22). The second step then computes modified plug-in bounds. I compute the plug-in bound with negative L^1 penalty on the λ, with $\delta \geq \delta^*$ being the penalty multiplier:

$$\hat{L}_{pen} = \max_{\lambda_1, \ldots, \lambda_K} \left[\frac{1}{N} \sum_{i=1}^{N} \min_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} - \delta \sum_{k=1}^{K} |\lambda_k| \right]. \tag{23}$$

I compute $\hat{U}_{pen}$ similarly with a positive L^1 penalty. The following proposition justifies use of the L^1 penalty:

Proposition 5. *Given the sample* $(W_1, \ldots, W_N)$ *and given* $\delta \in \mathbb{R}$*, consider the linear programming problem:*

$$\min_{P \in \mathcal{M}_{W \times V},\ P \geq 0} \int m(W_i, V_i) dP \qquad \text{subject to} \qquad \begin{aligned} &\left| \int \phi_k(W_i, V_i) dP \right| \leq \delta, \quad k = 1, \ldots, K, \\ &\int P(w, dv) = \hat{P}_W(w) \ \text{ for all } w \in \mathcal{W}. \end{aligned} \tag{24}$$

where $\hat{P}_W$ *is the empirical distribution of* W_i *constructed from* $(W_1, \ldots, W_N)$*. Its solution then equals* $\hat{L}_{pen}$*, defined in (23).*

Proof. See Appendix A.7. □

Proposition 5 shows that (23) equals the smallest value of θ for the distributions whose absolute-value GMM criterion of (20) is at most δ. In principle, such a distribution is not necessarily unique even when $\delta = \delta^*$. If it is unique, the resulting estimate of the identified set from the two-step procedure becomes a point.

In practice, due to either machine precision or the stopping criterion of numerical optimization methods, the numerical solution to (22) might be strictly smaller than the analytical solution δ^*. In that case, setting $\delta = \delta^*$ yields empty plug-in bound because the penalty multiplier is not large enough. To resolve this problem, a researcher may choose δ to be a little larger than δ^*, in which case (23) picks up the smallest value of θ for the distributions that attain the *near-minimum* of the absolute-value GMM criterion. In the special case that the minimizer distribution is unique, the resulting estimate of the identified set with $\delta > \delta^*$ is a very small interval instead of a point.

Although (23) resolves empty plug-in bound problem, it has two drawbacks. First, it is an ad-hoc approach and there is no formal justification for why relaxation of moment conditions is a good idea. Second, collection of distributions that minimize the criterion may yield a point (or a very small interval) identified set for the parameter of interest even if the model is partially identified. The literature dealt with the second problem by choosing δ that is reasonably larger than δ^* (Mogstad, Santos, and Torgovitsky, 2018), but how much larger it should be remains a question. The next subsection discusses a more principled approach, which is directly computing a confidence interval for the identified set.

5.4 Inference under empty plug-in bound

This subsection discusses a rigorous approach of dealing with empty plug-in bound, which is to propose a valid inference procedure under empty plug-in bound. The idea can be understood again by comparing it to GMM estimation. In GMM estimation, the point estimate is used as a building block for confidence interval, which assumes that the model is correctly specified. I take the same approach, using (23) as a building block for the inference procedure introduced earlier.

Recall that I proposed an inference procedure that tests (17) using the procedure of Chernozhukov, Chetverikov, and Kato (2019), which do not involve plug-in bound. The plug-in bound is involved only in the step of choosing Λ_F, which I propose to be a set of λs that are close to the solutions to plug-in bound problems. The inference procedure is valid regardless of whether the plug-in bound is empty or not; the issue is that there is no guidance for choosing Λ_F when plug-in bound is empty. In what follows, I propose guidance for how to choose Λ_F when plug-in bound is empty.

I propose using the previous subsection as a building block for choosing Λ_F. The procedure consists of three steps. The first step solves (22) and finds minimum δ^*. In the second step, consider a grid of positive real numbers $\{\delta_1, \ldots, \delta_M\}$ such that $\delta_m \geq \delta^*$ for all $m \in \{1, \ldots, M\}$. Then, for each δ_m, compute penalized plug-in bounds:

$$\begin{aligned}
\tilde{\lambda}_L(\delta_m) &= \underset{\lambda_1,\ldots,\lambda_K}{\operatorname{argmax}} \left[\frac{1}{N} \sum_{i=1}^{N} \min_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} - \delta_m \sum_{k=1}^{K} |\lambda_k| \right]. \\
\tilde{\lambda}_U(\delta m) &= \underset{\lambda_1,\ldots,\lambda_K}{\operatorname{argmin}} \left[\frac{1}{N} \sum_{i=1}^{N} \max_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} + \delta_m \sum_{k=1}^{K} |\lambda_k| \right].
\end{aligned} \tag{25}$$

Then I propose to sample points in the neighborhood of *each* $\tilde{\lambda}_L(\delta_m)$ and $\tilde{\lambda}_U(\delta_m)$. In simulation and application, I sample points by adding Gaussian noise to each $\tilde{\lambda}_L(\delta_m)$ and $\tilde{\lambda}_U(\delta_m)$ whose standard deviations are inversely proportional to their gradient at the solution to (25). I check performance of this approach by simulation in later section.

When $\delta^* = 0$, i.e. the plug-in bound is non-empty, a researcher can choose $\delta_1 = 0$ with $M = 1$. In this case, the inference procedure reduces to the procedure in Section 5.2, that is, the procedure under non-empty plug-in bound. This means that this subsection's inference procedure is a generalization of the procedure in Section 5.2.

6 Computation

This section discusses computation of objects discussed in the previous subsection which are not obvious to compute. The discussion focuses on the plug-in lower bound in (12), and the same discussion applies to all the other objects such as plug-in upper bound in (13), moment inequalities in (17) and penalized plug-in bounds in (25).

I present results regarding computation in two subsections. The first shows that for random coefficient models, the inner problem can be solved by a fast and exact algorithm for global polynomial optimization. The second shows that the outer problem is a convex optimization problem and hence easy to solve, given that the inner problems are solved fast and exactly.

6.1 The inner problem

The inner optimization problem of (12) is to evaluate the function

$$G_L(\lambda, w) = \min_{v \in \mathcal{V}} \left\{ m(w, v) + \sum_{k=1}^{K} \lambda_k \phi_k(w, v) \right\} \tag{26}$$

for each fixed $w = W_i$, where $i = 1, \ldots, N$, given the value of $\lambda \in \mathbb{R}^K$.

One difficulty in evaluating G_L is that the minimization problem must be solved globally. In the simple case that $\mathcal{V}$ is discrete or low-dimensional such as $\mathbb{R}$ or $\mathbb{R}^2$, the inner problem can be solved by enumerating all points in $\mathcal{V}$ or the grid points of $\mathcal{V}$. However, for random coefficient models, it is difficult to justify that the random coefficients are discrete. The dimension of $\mathcal{V}$ is also often large, because the dimension equals the number of regressors, including a constant.

This subsection shows that G_L can be computed fast and exactly when m and ϕ_ks are polynomials in v. When m and the ϕ_ks are polynomials, evaluation of G_L is equivalent to globally minimizing a polynomial, for which a fast and exact algorithm exists. The polynomial case appears when computing bounds for many interesting parameters, such as the moments and CDFs of random coefficients. The following examples describe some of them.

Example 5. Consider identification of the mean parameter $\mathbb{E}(e'V_i)$ discussed in Section 3. Theorem 1 characterized the identified set of $\mu_e = \mathbb{E}(e'V_i)$ under Assumptions 1 and 3. In this setup, the m function is given by

$$m(W_i, V_i) = e'V_i$$

which is a linear function of V_i and hence a first-order polynomial. The ϕ_ks under Assumption 3 consist of the functions

$$(R_{it}'V_i)(Y_{it} - R_{it}'V_i), \quad t = 1, \ldots, T, \tag{27}$$

and the entries of the vectors

$$(Z_i', X_i^{t\prime})'(Y_i - R_{it}'V_i), \quad t = 1, \ldots, T, \tag{28}$$

which are at most second-order polynomials of V_i. These moment restrictions are what I use in the application when estimating identified sets of the means of random coefficients.

Example 6. Suppose a researcher is interested in identifying an element of $\mathbb{E}(V_iV_i')$. Then m is an element of V_iV_i', which is a second-order polynomial of V_i. Suppose that the researcher

assumes the moment condition $\mathbb{E}((R_i'V_i)^3\varepsilon_{it}) = 0$, in which case the ϕ_ks consist of the functions

$$(R_{it}'V_i)^3(Y_{it} - R_{it}'V_i), \quad t = 1, \ldots, T,$$

which are fourth-order polynomials of V_i. The researcher might also assume that Assumption 3 holds, in which case he/she sets the additional ϕ_ks to be (27) and (28). These moment restrictions are what I use in the application when estimating identified sets of the variances and correlations of random coefficients.

In Examples 5 and 6, the moment functions are chosen so that they yield finite lower and upper bounds for the parameters of interest. As a practical strategy to ensure finite bounds, a researcher can choose ϕ_ks so that the inner objective function is an even order polynomial whose order is strictly larger than the order of the parameter of interest. In Examples 5 and 6, I choose ϕ_ks to be a second order polynomial for $\mathbb{E}(V_i)$ and a fourth order for $\mathbb{E}(V_iV_i')$. The inner objective function then has its leading coefficient positive or negative, depending on the signs of λ, which yields finite inner solutions in (12) and (13).

The polynomial case can be extended to allow either m or ϕ_ks to be indicator functions of V_i. The idea is that an indicator function partitions $\mathcal{V}$ into two exclusive sets, and the indicator function is constant within each set. A researcher can then compute the global optimum in each partition, and then the optimum of the two.

This extension is useful when computing bounds for CDFs of random coefficients, which is described in the following example.

Example 7. Let V_{i1} be the first entry of $V_i \in \mathbb{R}^{q+p}$, and let $v^0 \in \mathbb{R}$. Suppose a researcher is interested in identifying the CDF of V_{i1} evaluated at v^0. The researcher sets m to be

$$m(W_i, V_i) = \mathbf{1}(V_{i1} \leq v^0),$$

which is an indicator function of V_i. Assume that Assumption 3 holds, in which case the ϕ_ks are at most second-order polynomials in V_i, as shown in Example 5. The m function partitions the $\mathcal{V}$ space into two exclusive sets $\mathcal{V}_1 = \{(v_1, \ldots, v_{q+p}) \mid v_1 \leq v\}$ and $\mathcal{V}_2 = \{(v_1, \ldots, v_{q+p}) \mid v_1 > v\}$, and $m = 1$ on $\mathcal{V}_1$ and $m = 0$ on $\mathcal{V}_2$. The objective function in (26) is then a second-order polynomial in each of $\mathcal{V}_1$ and $\mathcal{V}_2$, for which the researcher can compute the minimum. The researcher can then evaluate G by taking the smaller optimum between those in $\mathcal{V}_1$ and $\mathcal{V}_2$.

The next two subsections discuss a fast and exact computation method for global optimization of polynomials. The first considers a simple case of quadratic polynomials for which the

global solution can be obtained in a closed-form. The second considers generic polynomials for which the global optimization problem is solved numerically.

6.1.1 Global optimization of quadratic polynomials

I first consider a simple case of global optimization of quadratic polynomials. I express a quadratic polynomial in standard form:

$$Q(v) = v'Av + b'v + c$$

where A is a $\dim(v) \times \dim(v)$ symmetric matrix, b is a $\dim(v)$-dimensional vector, and $c \in \mathbb{R}$. If the objective of (26) is expressed in this standard form, (A, b, c) are functions of w.

Quadratic polynomials can be solved efficiently using quadratic optimization softwares. In practice, a researcher can use a heuristic but faster (that is, closed-form) method. The first and second derivatives of $Q(v)$ are:

$$\frac{dQ}{dv} = 2Av + b, \qquad \frac{d^2Q}{dvdv'} = 2A.$$

If A is positive definite, Q is globally convex and has a global finite minimum at the solution to the first-order condition

$$\frac{dQ}{dv} = 2Av + b = 0$$

whose unique solution is $v^* = -(1/2)A^{-1}b$. Thus, the global minimum of Q is:

$$\min_{v \in \mathcal{V}} Q(v) = c - \frac{1}{4}b'A^{-1}b. \tag{29}$$

If A is not positive definite, A has a non-positive eigenvalue. If A has a negative eigenvalue, the minimum of Q is negative infinity. If A does not have a negative eigenvalue, which means A has a zero eigenvalue, A is singular and the only case in which Q has a finite minimum is when the first-order condition

$$2Av + b = 0$$

has an infinite number of solutions. If this is the case and the value of Q is constant over the solutions, Q has a finite minimum in any of the solutions. Otherwise, Q does not have a finite global minimum.

In practice, when solving (26) for each $w = W_i$ and if W_i follows a continuous distribution, A has a zero eigenvalue with probability zero. Therefore, a researcher may simply assign

negative infinity whenever A is not positive definite. That is, a researcher may simply use (29) to express (26) in a closed-form if and only if A is positive definite; otherwise the solution is negative infinity.

6.1.2 Global optimization of generic polynomials

When m and ϕ_ks are polynomials of generic order, a closed-form solution is unavailable, but it can be solved numerically. The idea is to transform the problem into a convex optimization problem (Lasserre, 2010, 2015). The resulting algorithm is fast and it computes an *exact* solution. This subsection discusses the main idea of the algorithm, and a formal discussion appears in the Appendix C.

Suppose a researcher wants to compute the global minimum of a fourth-order polynomial in two variables (v_1, v_2). Let $u(v) = (1, v_1, v_2, v_1^2, v_1 v_2, v_2^2)'$ be the vector of monomials up to the second order and $u_j(v)$ be the j-th entry of $u(v)$. Let $\{p_j(v)\}$ be the collection of all monomials up to the fourth order, which are unique entries of $\text{vec}(u(v)u(v)')$. Let J be the cardinality of $\{p_j(v)\}$.

Let a_j be the coefficient on the monomial $p_j(v)$. I can express a fourth-order polynomial in standard form:

$$\pi(v) = \sum_{j=1}^{J} a_j p_j(v).$$

Consider minimization of $\pi(v)$ with respect to $v \in \mathcal{V}$. The minimum of $\pi(v)$ over $\mathcal{V}$ equals the solution of minimization problem:

$$\min_{P_V \in \mathcal{M}_V,\ \int dP_V = 1} \int \pi(v) dP_V \tag{30}$$

where P_V is a probability distribution on $\mathcal{V}$. (30) is minimized at the point-mass distribution, concentrated at the minimizer of $\pi(v)$.

Since $\pi(v)$ is a linear combination of $p_j(v)$, I can rewrite (30) as:

$$\min_{P_V \in \mathcal{M}_V,\ \int dP_V = 1} \sum_{j=1}^{J} a_j \int p_j(v) dP_V,$$

which can be rewritten further as:

$$\min_{M_1,\ldots,M_J \in \mathbb{R},\ M_1 = 1} \sum_{j=1}^{J} a_j M_j \qquad \text{subject to} \qquad M_j = \int p_j(v) dP_V \text{ for some } P_V \in \mathcal{M}_V. \tag{31}$$

Except for the fact that the constraint is complicated, (31) is a minimization over $\mathbb{R}^J$ and the objective is linear, and hence convex, in the choice variables.

The idea is to replace the constraint in (31) with a convex constraint that only involves $(M_1, \ldots, M_J)$. The constraint in (31) indicates that $(M_1, \ldots, M_J)$ must be moments of some underlying distribution. Checking this constraint relates to a problem called *the moment problem* in mathematics; "Given the sequence of real numbers $(M_1, \ldots, M_J)$, can they be justified as moments of some distribution?"

A sequence of real numbers must satisfy some relationship between them for them to be justified as moments. For example, for a generic real random variable X, it must be that $\text{Var}(X)$ is positive. That is:

$$\mathbb{E}(X^2) - \mathbb{E}(X)^2 \geq 0.$$

This is equivalent to condition:

$$\begin{pmatrix} 1 & \mathbb{E}(X) \\ \mathbb{E}(X) & \mathbb{E}(X^2) \end{pmatrix} \text{ is positive semidefinite.}$$

This simple example can be generalized. Define linear operator $\mathcal{L}$ that maps a polynomial to $\mathbb{R}$ by relationship:

$$\mathcal{L}\left(\sum_j a_j p_j(v)\right) = \sum_j a_j M_j.$$

If $(M_1, \ldots, M_J)$ are moments, then

$$\mathcal{L}(u(v)u(v)') \text{ is positive semidefinite} \tag{32}$$

where the operator $\mathcal{L}$ is applied to each element of $u(v)u(v)'$. $\mathcal{L}(u(v)u(v)')$ is a matrix that involves only $(M_1, \ldots, M_J)$.

(32) is a convex constraint, based on the fact that the set of positive semidefinite matrices is a convex set in the space of vectorized matrix entries. Therefore, if I replace the constraint in (31) with (32), I obtain optimization problem:

$$\min_{M_1,\ldots,M_J \in \mathbb{R}} \sum_{j=1}^{J} a_j M_j \qquad \text{subject to} \qquad \mathcal{L}(u(v)u(v)') \text{ is positive semidefinite.} \tag{33}$$

The constraint can be handled more efficiently than a generic convex constraint so that the optimization problem has its own name—semidefinite program (SDP)—an optimization problem in which a matrix that involves the choice variables is constrained to be positive semidef-

inite.

The SDP approach to polynomial optimization solves (33), the *semidefinite relaxation*, which can be solved fast and reliably using SDP solvers available in the industry. The algorithm offers *certificate of optimality*, a condition for the optimal value of $(M_1, \ldots, M_J)$, satisfying which means that the solution to (33) equals the global optimum. For researchers interested in using the semidefinite relaxation approach to global polynomial optimization, I offer a general-purpose R package optpoly that implements the approach[8]. Alternatively, a general-purpose package, Gloptipoly (Henrion, Lasserre, and Löfberg, 2008), is available for Matlab users.

Since a necessary condition is weaker than the original condition, the solution to (33) (i.e., the SDP solution) is less than or equal to the solution to (31). The semidefinite relaxation approach solves a hierarchy of the SDP programs, or a *sequence* of the SDP programs, until the certificate of optimality is obtained, which is known to be obtained in a finite number of steps under suitable conditions. Even if a researcher does not solve the hierarchy of the SDPs, he/she can take an SDP solution as a lower bound for (31), and the resulting value of (12) is a conservative and yet valid lower bound for θ.

Instead of the SDPs, a researcher may solve a hierarchy of linear programs (LP) — the *LP relaxations* — for the global polynomial optimization problem (Lasserre, 2010, 2015). The LP hierarchy does not generally converge in finite steps and hence only asymptotic, but it can handle larger scale problems than the SDP hierarchy. Recently, Lasserre, Toh, and Yang (2017) and Weisser, Lasserre, and Toh (2018) proposed a relaxation that combines ideas of the SDP and LP hierarchies. Gautier and Rose (2019) used the latter, which exploits sparsity, in the context of instrumental variables models.

6.2 The outer problem

I turn to the outer optimization problem of (12). A researcher needs to solve the optimization problem:

$$\max_{\lambda \in \mathbb{R}^K} \frac{1}{N} \sum_{i=1}^{N} G_L(\lambda, W_i).$$

Assume that the researcher can evaluate G_L exactly using the algorithm in the previous subsection. The remaining difficulty then is how to solve the optimization problem given that K is potentially large. The following proposition shows that the outer optimization problem is a convex optimization problem.

[8] Available at `https://github.com/wooyong/optpoly`.

Proposition 6. *Define*

$$\hat{L}(\lambda) = \frac{1}{N}\sum_{i=1}^{N} G_L(\lambda, W_i)$$

where G_L is defined in (12). $G_L(\lambda, W_i)$ is globally concave in λ, which implies global concavity of $\hat{L}(\lambda)$.

Proof. See Appendix A.8. □

Proposition 6 suggests that there is only one local maximum of $\hat{L}(\lambda)$, which is also the global maximum. Milgrom and Segal (2002, Theorem 3) provides conditions under which G_L is differentiable when $K = 1$, which can be used to provide conditions under which G_L is directionally differentiable. This suggests that the researcher can maximize $\hat{L}(\lambda)$ using fast convex optimization algorithms such as gradient descent methods. In practice, if a researcher is concerned with differentiability, he/she can apply gradient descent methods based on finite differences.

Concavity of $\hat{L}(\lambda)$ comes from the concavity of G_L, and solving the inner problem exactly by the polynomial optimization algorithm is crucial to computational tractability of the outer problem. This is an important distinction from the general-purpose approaches of Schennach (2014) and Li (2018). I focus on random coefficient models and exploit the structure of the model to achieve computational tractability for the models with large dimensions. If a researcher uses general-purpose global optimization methods such as simulated annealing to solve the inner problem, then G is no longer concave and the researcher cannot use fast convex optimization algorithms for the outer problem. This is problematic when K is large, which is often the case in random coefficient models. For example, during application of Section 8, K can be as large as several hundreds.

7 Simulation

This section performs simulation regarding inference procedure discussed in the previous sections. All simulations consider the AR(1) model given in (3):

$$Y_{it} = \gamma_i + \beta_i Y_{i,t-1} + \varepsilon_{it}, \quad t = 1, \ldots, T.$$

$\gamma_i \in \mathbb{R}$ and $\beta_i \in [0, 1]$ follow Normal and Beta distributions respectively, and their joint distribution is given by Gaussian copula. ε_{it} follows an independent Normal distribution with mean zero and variance varying over t. Y_{i0} is generated from the stationary distribution

implied by (γ_i, β_i) and ε_{i1} when $\beta_i \leq 0.9$:

$$Y_{i0} \sim N\left(\frac{\gamma_i}{1-\beta_i}, \frac{\text{Var}(\varepsilon_{i1})}{1-\beta_i^2}\right).$$

When $\beta_i > 0.9$, the stationary distribution produces extreme values because of small denominator values, so Y_{i0} is generated from independent Normal distribution whose mean and variance match the distribution of Y_{i0} given $\beta_i \leq 0.9$. Parameter values for the distributions of $(\gamma_i, \beta_i, \varepsilon_{it})$ are chosen to resemble the estimates of income process in the application.

Simulation data are generated in two steps. The first step simulates a dataset of 100,000 observations from the parametric model described above. The second step then creates Monte Carlo samples from the 100,000 observations, by sampling observations with replacement. This means that the 100,000 observations of the first step is a *finite population* from which Monte Carlo samples are generated in the second step. Having a finite population is convenient because I can exactly compute the population identified set using (12) and (13), whereas computing the identified set of the parametric model itself is infeasible.

Table 1 lists bounds for $\mathbb{E}(\beta_i)$ computed from the finite population with $T = 5, 10, 15$. For each T, I compute the bounds using moment conditions:

$$\begin{aligned}
\mathbb{E}((\gamma_i + \beta_i Y_{i,t-1})\varepsilon_{it}) &= 0, \quad t = 1, \ldots, T, \\
\mathbb{E}(\varepsilon_{it}) &= 0, \quad t = 1, \ldots, T, \\
\mathbb{E}(Y_{i,t-1-s}\varepsilon_{it}) &= 0, \quad s = 0, \ldots, \min\{L, T\}, \quad t = 1, \ldots, T,
\end{aligned}$$

where $L = 3, 5$ or 7. I also restrict $(\gamma_i, \beta_i) \in \mathcal{V} = [-3, 3] \times [0, 1]$ during computation, which is true for finite population used in the simulation. I use the same setup during application with $T = 15$ and $L = 7$.

	$T = 5$	$T = 10$	$T = 15$
$L = 3$	[0.195, 0.834]	[0.317, 0.738]	[0.371, 0.685]
$L = 5$	[0.198, 0.825]	[0.319, 0.720]	[0.372, 0.675]
$L = 7$	[0.196, 0.823]	[0.319, 0.728]	[0.368, 0.669]

Table 1: Bounds for $\mathbb{E}(\beta_i)$ computed from the 100,000 observations from parametric model, for each T and L. I use the 100,000 observations as finite population for which the above is population bound for $\mathbb{E}(\beta_i)$.

For each (T, L), I create Monte Carlo replications by sampling $N = 750$ or 1000 observations from the finite population with replacement. I then compute confidence interval for $\mathbb{E}(\beta_i)$ for each Monte Carlo replication, using generalized inference procedure discussed in

Section 5.4. The grid of $\{\delta_m\}$ is set to be

$$\delta_m \in \{1.25\delta^*, 1.5\delta^*, 1.75\delta^*, 2.0\delta^*, 2.25\delta^*, 2.5\delta^*, 2.75\delta^*, 3\delta^*\}.$$

For each $\tilde{\lambda}_L(\delta_m)$ and each $\tilde{(}\lambda_U(\delta_m))$, I sample $P = 25, 50$ or 75 points from their neighborhood by adding Gaussian noise whose standard deviation is inversely proportional to the gradient of λ at the $\tilde{\lambda}$s. This means that the size of Λ_F — the number of moment inequalities — is $8P$. The critical value is computed with 2000 multiplier bootstrap replications.

Tables 2 and 3 present coverage probabilities for combinations of N, T, L, P. Each coverage probability is computed with 1000 Monte Carlo replications. Simulation results suggest that the proposed inference procedure produces conservative but reasonable coverage probabilities.

$T = 5$	$L = 3$	$L = 5$	$L = 7$
$P = 25$	0.956	0.951	0.946
$P = 50$	0.971	0.963	0.959
$P = 75$	0.974	0.966	0.971
$T = 10$	$L = 3$	$L = 5$	$L = 7$
$P = 25$	0.933	0.929	0.894
$P = 50$	0.949	0.955	0.919
$P = 75$	0.960	0.962	0.937
$T = 15$	$L = 3$	$L = 5$	$L = 7$
$P = 25$	0.986	0.929	0.874
$P = 50$	0.990	0.947	0.915
$P = 75$	0.991	0.961	0.939

Table 2: Coverage probabilities of inference procedures for $N = 750$ sample size, with nominal coverage probability of 0.9.

8 Application to lifecycle earnings dynamics

8.1 Overview

Lifecycle earnings dynamics is a key input in various macroeconomic models. For example, in models of consumption and savings dynamics (Hall and Mishkin, 1982; Blundell, Pistaferri, and Preston, 2008; Blundell, Pistaferri, and Saporta-Eksten, 2016; Arellano, Blundell, and Bonhomme, 2017), households with higher risk in earnings dynamics accumulate more precautionary savings to smooth consumption. Households with persistent earnings shocks

$T = 5$	$L = 3$	$L = 5$	$L = 7$
$P = 25$	0.959	0.943	0.949
$P = 50$	0.968	0.957	0.960
$P = 75$	0.972	0.964	0.966
$T = 10$	$L = 3$	$L = 5$	$L = 7$
$P = 25$	0.869	0.890	0.885
$P = 50$	0.923	0.916	0.925
$P = 75$	0.934	0.930	0.938
$T = 15$	$L = 3$	$L = 5$	$L = 7$
$P = 25$	0.974	0.880	0.821
$P = 50$	0.988	0.910	0.869
$P = 75$	0.990	0.933	0.904

Table 3: Coverage probabilities of inference procedures for $N = 1000$ sample size, with nominal coverage probability of 0.9.

save a lot when they experience a persistent positive earnings shock, which is used to maintain consumption during a persistent negative earnings shock. Specifying earnings process that highlights features of real data is important for calibrating and drawing conclusions from these models.

When used as an input, it is common to specify earnings dynamics using a parsimonious linear model. It consists of permanent and transitory income processes:

$$Y_{it} = z_{it} + \varepsilon_{it}, \quad z_{it} = \rho z_{i,t-1} + \eta_{it},$$

where Y_{it} is residual log-earnings, i.e. log-earnings net of common trends on observables such as demographics and years of experience, $\{z_{it}\}$ is permanent income process and $\{\varepsilon_{it}\}$ is transitory income process. η_{it} and ε_{it} are i.i.d. mean zero shocks.

The literature has two leading views on unobserved heterogeneity in earnings dynamics. Consider two residual earnings processes[9]:

$$\begin{aligned} Y_{it} &= \alpha_i + z_{it} + \varepsilon_{it}, & z_{it} &= \rho z_{i,t-1} + \eta_{it}, \\ Y_{it} &= \alpha_i + \beta_i h_{it} + z_{it} + \varepsilon_{it}, & z_{it} &= \rho z_{i,t-1} + \eta_{it}, \end{aligned} \tag{34}$$

where h_{it} is potential years of experience and (α_i, β_i) are heterogeneous deviations from common trends. The two models are called Restricted Income Profiles (RIP) process and Hetero-

[9]As Guvenen (2007) points out, these are stylized versions of what is used in the literature, but they still capture features important for the discussion.

geneous Income Profiles (HIP) process, respectively. Estimates of the second model (HIP) in the literature report $0.5 < \rho < 0.7$ and $\text{Var}(\beta_i) > 0$ (e.g. Lillard and Weiss, 1979; Baker, 1997), which means households experience modest persistence and heterogeneous trends. However, a seminal paper by MaCurdy (1982) tested hypothesis that $\text{Var}(\beta_i) = 0$ and did not reject the hypothesis. Estimates of the first model (RIP) in the literature report $\rho \approx 1$ (e.g. Abowd and Card, 1989; Topel and Ward, 1992), meaning households experience extreme persistence and homogeneous trends. Guvenen (2007) studied implications of the two models on consumption data and found that HIP is more consistent with features of consumption data. Guvenen (2009) pointed out that misspecifying HIP process as a RIP process leads to an upward biased estimator of ρ, therefore obtaining $\rho \approx 1$.

In contrast to vast literature on investigating heterogeneity in β_i, there is relatively little on investigating heterogeneity in ρ itself although ρ is a key parameter that affects household decisions. Recent studies include Browning, Ejrnaes, and Alvarez (2010) and Alan, Browning, and Ejrnæs (2018) who estimated a parametric model of income processes in which heterogeneity in ρ is given by a factor structure. In this section, I empirically investigate heterogeneity in ρ by estimating (34) with $\rho = \rho_i$. I treat (34) with $\rho = \rho_i$ as a random coefficient model, meaning that distribution of ρ_i and its dependence to (α_i, β_i) and initial earnings Y_{i0} are unrestricted. Distribution of η_{it}s are also not restricted and may depend on ρ_i, allowing for heteroskedasticity.

I find that, when $\rho = \rho_i$ is allowed to be heterogeneous, both processes have similar estimates of $\mathbb{E}(\rho_i)$ that is significantly less than 1. Confidence intervals for $\mathbb{E}(\rho_i)$ in the two processes overlap, having upper confidence limits around 0.6 at 90% confidence level. Confidence intervals for the CDF of ρ_i, i.e. $\mathbb{P}(\rho_i \leq r)$ for a grid of r, are also similar between the two processes. These results suggest that choosing HIP or RIP may not lead to serious misspecification when ρ is allowed to be heterogeneous. I also find heterogeneity in ρ_i by computing confidence interval for $\text{Var}(\rho_i)$ in the first model. It has a lower confidence limit of 0.02 at 90% confidence level, implying a lower confidence limit of 0.16 for standard deviation of ρ_i.

8.2 Data

I use data on U.S. households from the Panel Study of Income Dynamics (PSID) dataset. I use the data provided by Guvenen (2009) as a supplementary material, who estimated HIP and RIP processes using PSID earnings data of male head of households collected annually from 1968 to 1993. The dataset contains male head of households who are not in the poverty (SEO) subsample and who consecutively reported positive hours (between 520 and 5110 hours a year) and earnings (between a preset minimum and maximum wage). I also follow Guvenen

(2009) and use potential experience as a measure of experience:

$$h = \text{age} - \max\{\text{years of schooling}, 12\} - 6,$$

Note that potential experience is a deterministic function of age.

The data provided by Guvenen (2009) is an unbalanced panel. To create a balanced panel, I collect individuals with 16 consecutive waves of data, which yields $N = 800$ and $T = 15$ taking the first wave as an initial value of earnings.

8.3 From income processes to random coefficient models

To apply my method to income processes, I transform two models in (34) to random coefficient models. I first use the empirical deconvolution method of Arellano and Bonhomme (2108) to separate ε_{it} from Y_{it}, obtaining observations of $\tilde{Y}_{it}$:

$$\begin{aligned} \tilde{Y}_{it} &= \alpha_i + z_{it}, & z_{it} &= \rho z_{i,t-1} + \eta_{it}, \\ \tilde{Y}_{it} &= \alpha_i + \beta_i h_{it} + z_{it}, & z_{it} &= \rho z_{i,t-1} + \eta_{it}. \end{aligned} \tag{35}$$

I then use quasi-differencing to transform each model to a random coefficient model. Quasi-differencing the first model yields:

$$\begin{aligned} \tilde{Y}_{it} &= \alpha_i(1-\rho_i) &&+\rho_i \tilde{Y}_{i,t-1} + \eta_{it} \\ &= \tilde{\alpha}_i &&+\rho_i \tilde{Y}_{i,t-1} + \eta_{it}. \end{aligned}$$

The last line is a standard random coefficient model. Similarly, quasi-differencing the second model yields:

$$\begin{aligned} \tilde{Y}_{it} &= \alpha_i(1-\rho_i) + \beta_i\rho_i &&+\beta_i(1-\rho_i)h_{it} &&+\rho_i \tilde{Y}_{i,t-1} + \eta_{it} \\ &= \tilde{\alpha}_i &&+\tilde{\beta}_i h_{it} &&+\rho_i \tilde{Y}_{i,t-1} + \eta_{it}, \end{aligned}$$

which is also a standard random coefficient model. Note that h_{it} is a deterministic function of age, which is strictly exogenous.

8.4 Estimation and inference

For each model, I compute confidence intervals for $\mathbb{E}(\rho_i)$, $\text{Var}(\rho_i)$ and $\mathbb{P}(\rho_i \leq r)$ for a grid of values $r = 0, 0.1, \ldots, 0.9, 1$. For $\mathbb{E}(\rho_i)$ and $\mathbb{P}(\rho_i \leq r)$, I use orthogonality conditions stated in

Example 5. In particular, I use, for the first model:

$$\begin{aligned}\mathbb{E}((\tilde{\alpha}_i + \rho_i \tilde{Y}_{i,t-1})\eta_{it}) &= 0,\\ \mathbb{E}(\eta_{it}) &= 0,\\ \mathbb{E}(\tilde{Y}_{i,t-1-s}\eta_{it}) &= 0, \quad s = 0, \ldots, 5,\end{aligned}$$

and for the second model:

$$\begin{aligned}\mathbb{E}((\tilde{\alpha}_i + \tilde{\beta}_i h_{it} + \rho_i \tilde{Y}_{i,t-1})\eta_{it}) &= 0,\\ \mathbb{E}(\eta_{it}) &= 0,\\ \mathbb{E}(\tilde{Y}_{i,t-1-s}\eta_{it}) &= 0, \quad s = 0, \ldots, 5,\\ \mathbb{E}(h_{i,t-s}\eta_{it}) &= 0, \quad s = -5, \ldots, -1, 0, 1, \ldots, 5.\end{aligned}$$

I use additional orthogonality conditions for computing confidence interval of $\mathrm{Var}(\rho_i)$. Additional orthogonality conditions for the first model are:

$$\begin{aligned}\mathbb{E}((\tilde{\alpha}_i + \rho_i \tilde{Y}_{i,t-1})^3\eta_{it}) &= 0,\\ \mathbb{E}(\tilde{\alpha}_i^k \eta_{it}) &= 0, \quad k = 1, 2,\\ \mathbb{E}(\rho_i^k \eta_{it}) &= 0, \quad k = 1, 2,\end{aligned}$$

and for the second model are:

$$\begin{aligned}\mathbb{E}((\tilde{\alpha}_i + \tilde{\beta}_i h_{it} + \rho_i \tilde{Y}_{i,t-1})^3\eta_{it}) &= 0,\\ \mathbb{E}(\tilde{\alpha}_i^k \eta_{it}) &= 0, \quad k = 1, 2,\\ \mathbb{E}(\tilde{\beta}_i^k \eta_{it}) &= 0, \quad k = 1, 2,\\ \mathbb{E}(\rho_i^k \eta_{it}) &= 0, \quad k = 1, 2.\end{aligned}$$

These additional conditions imply finite lower and upper bounds on the second moments of $(\tilde{\alpha}_i, \tilde{\beta}_i, \rho_i)$.

With these orthogonality conditions, I compute confidence intervals using the procedure in Section 5.4. Tuning parameters for inference procedure are the same as those in simulations: the grid of δ is set to be

$$\delta_m \in \{1.25\delta^*, 1.5\delta^*, 1.75\delta^*, 2.0\delta^*, 2.25\delta^*, 2.5\delta^*, 2.75\delta^*, 3\delta^*\},$$

$P = 50$ points in the neighborhood of each $\tilde{\lambda}_L(\delta_m)$ and $\tilde{(}\lambda_U(\delta_m))$ are sampled using additive

	Confidence interval of $\mathbb{E}(\rho_i)$	Confidence interval of $\mathrm{Var}(\rho_i)$
First model	[0.513, 0.572]	[0.030, 0.335]
Second model	[0.222, 0.660]	[0.000, 0.676]

Table 4: Confidence interval for $\mathbb{E}(\rho_i)$ and $\mathrm{Var}(\rho_i)$. Nominal coverage probability is 0.9.

$\mathbb{P}(\rho_i \leq r)$	First model	Second model
$r = 0.0$	[0.000, 0.327]	[0.000, 0.723]
$r = 0.1$	[0.017, 0.333]	[0.005, 0.785]
$r = 0.2$	[0.118, 0.375]	[0.083, 0.857]
$r = 0.3$	[0.082, 0.539]	[0.075, 0.858]
$r = 0.4$	[0.145, 0.620]	[0.146, 0.901]
$r = 0.5$	[0.217, 0.725]	[0.174, 0.959]
$r = 0.6$	[0.280, 0.905]	[0.201, 0.983]
$r = 0.7$	[0.342, 0.943]	[0.243, 1.000]
$r = 0.8$	[0.463, 0.958]	[0.305, 1.000]
$r = 0.9$	[0.550, 0.984]	[0.355, 1.000]
$r = 1.0$	[0.608, 1.000]	[0.386, 1.000]

Table 5: Confidence intervals for $\mathbb{P}(\rho_i \leq r)$. Nominal coverage probability is 0.9.

Gaussian noise whose standard deviation is inversely proportional to the gradient of λ at the penalized plug-in bounds, and the critical value is computed with 2000 multiplier bootstrap replications.

8.5 Empirical results

Confidence intervals for $\mathbb{E}(\rho_i)$ and $\mathrm{Var}(\rho_i)$ are given in Table 4. Both models estimate $\mathbb{E}(\rho_i)$ to be significantly less than 1. Moreover, their confidence intervals show substantial overlap, having similar upper confidence limits of 0.572 in the first model and 0.660 in the second. These suggest that specifying homogeneous or heterogeneous β does not lead to serious misspecification when ρ is allowed to be heterogeneous. Confidence interval for $\mathrm{Var}(\rho_i)$ suggest heterogeneity in ρ_i, having a nonzero lower confidence limit of 0.030 for the first model implying standard deviation of 0.173.

Confidence intervals for the CDF of ρ_i for a grid of values are given in Table 5. They suggest substantial heterogeneity in ρ_i. For example, the first model estimates that at least 21.7% of households are estimated to have $\rho_i \leq 0.5$, while at least 27.5% of households are estimated to have $\rho_i > 0.5$. Confidence intervals for the two CDFs show substantial overlap as well.

9 Conclusion

This paper studies identification and estimation of dynamic random coefficient models. I show that the model is not point-identified, and I characterize a sharp identified set using the duality representation of infinite-dimensional linear programming. A computationally feasible estimation and inference procedure for the identified set is proposed, which uses a fast and exact algorithm for global polynomial optimization—the semidefinite relaxations approach. Inference of the identified set uses results from literature on moment inequalities models.

I estimate unobserved heterogeneity in earnings persistence across U.S. households using the PSID dataset. I find that the average earnings persistence is significantly less than 1 when it is allowed to be heterogeneous. I also find that, when earnings persistence is allowed to be heterogeneous, choosing RIP over HIP or vice versa may not lead to serious misspecification of the persistence. Estimates for variance and CDF of earnings persistence suggest that there is substantial degree of unobserved heterogeneity in it.

References

Abowd, John M and David Card. 1989. "On the covariance structure of earnings and hours changes." *Econometrica* 57 (2):411–445.

Ackerberg, Daniel A, Kevin Caves, and Garth Frazer. 2015. "Identification properties of recent production function estimators." *Econometrica* 83 (6):2411–2451.

Alan, Sule, Martin Browning, and Mette Ejrnæs. 2018. "Income and consumption: A micro semistructural analysis with pervasive heterogeneity." *Journal of Political Economy* 126 (5):1827–1864.

Anderson, Edward J. 1983. "A review of duality theory for linear programming over topological vector spaces." *Journal of Mathematical Analysis and Applications* 97 (2):380–392.

Andrews, Donald WK and Soonwoo Kwon. 2019. "Inference in moment inequality models that is robust to spurious precision under model misspecification." *Working paper* .

Andrews, Donald WK and Xiaoxia Shi. 2013. "Inference based on conditional moment inequalities." *Econometrica* 81 (2):609–666.

Arellano, Manuel, Richard Blundell, and Stéphane Bonhomme. 2017. "Earnings and consumption dynamics: a nonlinear panel data framework." *Econometrica* 85 (3):693–734.

Arellano, Manuel and Stephen Bond. 1991. "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations." *Review of Economic Studies* 58 (2):277–297.

Arellano, Manuel and Stéphane Bonhomme. 2012. "Identifying distributional characteristics in random coefficients panel data models." *Review of Economic Studies* 79 (3):987–1020.

———. 2108. "Recovering latent variables by matching." *Working paper* .

Bai, Yuehao, Andres Santos, and Azeem M Shaikh. 2019. "A practical method for testing many moment inequalities." *Working Paper* .

Baker, Michael. 1997. "Growth-rate heterogeneity and the covariance structure of life-cycle earnings." *Journal of Labor Economics* 15 (2):338–375.

Bierens, Herman J. 1990. "A consistent conditional moment test of functional form." *Econometrica: Journal of the Econometric Society* :1443–1458.

Blundell, Richard and Stephen Bond. 1998. "Initial conditions and moment restrictions in dynamic panel data models." *Journal of Econometrics* 87 (1):115–143.

Blundell, Richard, Luigi Pistaferri, and Ian Preston. 2008. "Consumption inequality and partial insurance." *American Economic Review* 98 (5):1887–1921.

Blundell, Richard, Luigi Pistaferri, and Itay Saporta-Eksten. 2016. "Consumption inequality and family labor supply." *American Economic Review* 106 (2):387–435.

Browning, Martin, Mette Ejrnaes, and Javier Alvarez. 2010. "Modelling income processes with lots of heterogeneity." *Review of Economic Studies* 77 (4):1353–1381.

Chamberlain, Gary. 1992. "Efficiency bounds for semiparametric regression." *Econometrica* 60 (3):567–596.

———. 1993. "Feedback in panel data models." *Working paper* .

Chernozhukov, Victor, Denis Chetverikov, and Kengo Kato. 2019. "Inference on causal and structural parameters using many moment inequalities." *Review of Economic Studies* 86 (5):1867–1900.

Chernozhukov, Victor, Sokbae Lee, and Adam M Rosen. 2013. "Intersection bounds: Estimation and inference." *Econometrica* 81 (2):667–737.

Chesher, Andrew and Adam M Rosen. 2017. "Generalized instrumental variable models." *Econometrica* 85 (3):959–989.

Galichon, Alfred and Marc Henry. 2009. "A test of non-identifying restrictions and confidence regions for partially identified parameters." *Journal of Econometrics* 152 (2):186–196.

———. 2011. "Set identification in models with multiple equilibria." *Review of Economic Studies* 78 (4):1264–1298.

Gautier, Eric and Christiern Rose. 2019. "High-dimensional instrumental variables regression and confidence sets." *Working paper, arXiv preprint arXiv:1105.2454* .

Graham, Bryan S and James L Powell. 2012. "Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models." *Econometrica* 80 (5):2105–2152.

Gunsilius, Florian. 2019. "Bounds in continuous instrumental variable models." *Working paper, arXiv preprint arXiv:1910.09502* .

Guvenen, Fatih. 2007. "Learning your earning: Are labor income shocks really very persistent?" *American Economic Review* 97 (3):687–712.

———. 2009. "An empirical investigation of labor income processes." *Review of Economic dynamics* 12 (1):58–79.

Hall, Robert E and Frederic S Mishkin. 1982. "The sensitivity of consumption to transitory income: Estimates from panel data on households." *Econometrica* 50 (2):461–481.

Henrion, Didier, Jean-Bernard Lasserre, and Johan Löfberg. 2008. *GloptiPoly 3: moments, optimization and semidefinite programming*.

Honoré, Bo E and Elie Tamer. 2006. "Bounds on parameters in panel dynamic discrete choice models." *Econometrica* 74 (3):611–629.

Jappelli, Tullio and Luigi Pistaferri. 2010. "The consumption response to income changes." *Annual Review of Economics* 2:479–506.

Kiefer, Jack. 1959. "Optimum experimental designs." *Journal of the Royal Statistical Society: Series B* 21 (2):272–304.

Lasserre, Jean B, Kim-Chuan Toh, and Shouguang Yang. 2017. "A bounded degree SOS hierarchy for polynomial optimization." *EURO Journal on Computational Optimization* 5 (1-2):87–117.

Lasserre, Jean-Bernard. 2010. *Moments, positive polynomials and their applications*. World Scientific.

———. 2015. *An introduction to polynomial and semi-algebraic optimization*. Cambridge University Press.

Levinsohn, James and Amil Petrin. 2003. "Estimating production functions using inputs to control for unobservables." *Review of Economic Studies* 70 (2):317–341.

Li, Lixiong. 2018. "Identification of structural and counterfactual parameters in a large class of structural econometric models." *Working paper* .

Lillard, Lee A and Yoram Weiss. 1979. "Components of variation in panel earnings data: American scientists 1960-70." *Econometrica* 47 (2):437–454.

MaCurdy, Thomas E. 1982. "The use of time series processes to model the error structure of earnings in a longitudinal data analysis." *Journal of Econometrics* 18 (1):83–114.

Milgrom, Paul and Ilya Segal. 2002. "Envelope theorems for arbitrary choice sets." *Econometrica* 70 (2):583–601.

Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2018. "Using instrumental variables for inference about policy relevant treatment parameters." *Econometrica* 86 (5):1589–1619.

Newey, Whitney K and Daniel McFadden. 1994. "Large sample estimation and hypothesis testing." *Handbook of Econometrics* 4:2111–2245.

Nie, Jiawang, James Demmel, and Bernd Sturmfels. 2006. "Minimizing polynomials via sum of squares over the gradient ideal." *Mathematical Programming* 106 (3):587–606.

Nordström, Kenneth. 2011. "Convexity of the inverse and Moore–Penrose inverse." *Linear algebra and its applications* 434 (6):1489–1512.

Olley, G Steven and Ariel Pakes. 1996. "The dynamics of productivity in the telecommunications equipment industry." *Econometrica* 64 (6):1263–1297.

Romano, Joseph P, Azeem M Shaikh, and Michael Wolf. 2014. "A practical two-step method for testing moment inequalities." *Econometrica* 82 (5):1979–2002.

Schennach, Susanne M. 2014. "Entropic latent variable integration via simulation." *Econometrica* 82 (1):345–385.

Stinchcombe, Maxwell B and Halbert White. 1998. "Consistent specification testing with nuisance parameters present only under the alternative." *Econometric Theory* 14 (3):295–325.

Topel, Robert H and Michael P Ward. 1992. "Job mobility and the careers of young men." *Quarterly Journal of Economics* 107 (2):439–479.

Torgovitsky, Alexander. 2019. "Nonparametric inference on state dependence in unemployment." *Working paper, Available at SSRN:2564305* .

Van der Vaart, Aad W. 2000. *Asymptotic statistics*, vol. 3. Cambridge university press.

Weisser, Tillmann, Jean B Lasserre, and Kim-Chuan Toh. 2018. "Sparse-BSOS: a bounded degree SOS hierarchy for large scale polynomial optimization with sparsity." *Mathematical Programming Computation* 10 (1):1–32.

Wooldridge, Jeffrey M. 2005. "Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models." *Review of Economics and Statistics* 87 (2):385–390.

Appendices

A Proofs

A.1 Proof of Proposition 1

For notational simplicity, assume $\mathcal{C} = \mathcal{C}_0^5$ where $\mathcal{C}_0$ is a compact subset of $\mathbb{R}$. The proof can be easily modified for a generic compact set $\mathcal{C}$.

Let $f : \mathcal{C}_0^3 \mapsto \mathbb{R}$, $g_1 : \mathcal{C}_0^3 \mapsto \mathbb{R}$ and $g_2 : \mathcal{C}_0^4 \mapsto \mathbb{R}$ be bounded functions with respect to the Lebesgue measure almost everywhere. From the proof of Theorem 3 in the Appendix, the sharp lower bound of $\mathbb{E}(\beta_i)$ equals:

$$\begin{aligned} &\max_{f,g_1,g_2} \mathbb{E}(f(Y_{i0}, Y_{i1}, Y_{i2})) \qquad \text{subject to} \\ &\qquad f(Y_{i0}, Y_{i1}, Y_{i2}) + g_1(\gamma_i, \beta_i, Y_{i0})\varepsilon_{i1} + g_2(\gamma_i, \beta_i, Y_{i0}, Y_{i1})\varepsilon_{i2} \leq \beta_i. \end{aligned} \tag{36}$$

The sharp upper bound of $\mathbb{E}(\beta_i)$ equals:

$$\begin{aligned} &\min_{f,g_1,g_2} \mathbb{E}(f(Y_{i0}, Y_{i1}, Y_{i2})) \qquad \text{subject to} \\ &\qquad f(Y_{i0}, Y_{i1}, Y_{i2}) + g_1(\gamma_i, \beta_i, Y_{i0})\varepsilon_{i1} + g_2(\gamma_i, \beta_i, Y_{i0}, Y_{i1})\varepsilon_{i2} \geq \beta_i. \end{aligned} \tag{37}$$

I will suppose $\mathbb{E}(\beta_i)$ is point-identified and derive contradiction. The argument uses the following proposition.

Proposition 7. *Suppose that $\mathbb{E}(\beta_i)$ is point-identified. Then there exists (f^*, g_1^*, g_2^*) such that*

$$f^*(Y_{i0}, Y_{i1}, Y_{i2}) + g_1^*(\gamma_i, \beta_i, Y_{i0})\varepsilon_{i1} + g_2^*(\gamma_i, \beta_i, Y_{i0}, Y_{i1})\varepsilon_{i2} = \beta_i \tag{38}$$

almost surely on $\mathcal{C}_0^5$.

Proof. Suppose such functions do not exist. Then the solution to (36), denoted by (f^l, g_1^l, g_2^l), satisfy

$$f^l(Y_{i0}, Y_{i1}, Y_{i2}) + g_1^l(\gamma_i, \beta_i, Y_{i0})\varepsilon_{i1} + g_2^l(\gamma_i, \beta_i, Y_{i0}, Y_{i1})\varepsilon_{i2} \leq \beta_i$$

where inequality is strict on a positive Lebesgue measure on $\mathcal{C}_0^5$. Similarly, the solution to (37), denoted by (f^u, g_1^u, g_2^u), satisfy

$$f^u(Y_{i0}, Y_{i1}, Y_{i2}) + g_1^u(\gamma_i, \beta_i, Y_{i0})\varepsilon_{i1} + g_2^u(\gamma_i, \beta_i, Y_{i0}, Y_{i1})\varepsilon_{i2} \geq \beta_i$$

where inequality is strict on a positive Lebesgue measure on $\mathcal{C}_0^5$. Then:

$$\begin{aligned}\mathbb{E}(f^l(Y_{i0}, Y_{i1}, Y_{i2})) &= \mathbb{E}\left(f^l(Y_{i0}, Y_{i1}, Y_{i2}) + g_1^l(\gamma_i, \beta_i, Y_{i0})\varepsilon_{i1} + g_2^l(\gamma_i, \beta_i, Y_{i0}, Y_{i1})\varepsilon_{i2}\right) \\ &< \mathbb{E}(\beta_i) \\ &< \mathbb{E}\left(f^u(Y_{i0}, Y_{i1}, Y_{i2}) + g_1^u(\gamma_i, \beta_i, Y_{i0})\varepsilon_{i1} + g_2^u(\gamma_i, \beta_i, Y_{i0}, Y_{i1})\varepsilon_{i2}\right) \\ &= \mathbb{E}(f^u(Y_{i0}, Y_{i1}, Y_{i2}))\end{aligned}$$

where strict inequalities follow because the density of $(\gamma_i, \beta_i, Y_{i0}, Y_{i1}, Y_{i2})$ has a lower bound $b > 0$. The above implies the sharp lower bound $\mathbb{E}(f^l(Y_{i0}, Y_{i1}, Y_{i2}))$ is strictly less than the sharp upper bound $\mathbb{E}(f^u(Y_{i0}, Y_{i1}, Y_{i2}))$. This is contradiction since $\mathbb{E}(\beta_i)$ is assumed to be point-identified. □

To derive contradiction, suppose $\mathbb{E}(\beta_i)$ is point-identified. Substitute $\varepsilon_{it} = Y_{it} - \gamma_i - \beta_i Y_{i,t-1}$ in (38) and obtain:

$$f^*(Y_{i0}, Y_{i1}, Y_{i2}) + g_1^*(\gamma_i, \beta_i, Y_{i0})(Y_{i1} - \gamma_i - \beta_i Y_{i0}) + g_2^*(\gamma_i, \beta_i, Y_{i0}, Y_{i1})(Y_{i2} - \gamma_i - \beta_i Y_{i1}) = \beta_i. \tag{39}$$

Take any $\gamma, \tilde{\gamma}, \beta, y_0, y_1, y_2 \in \mathcal{C}_0$ such that $\gamma \neq \tilde{\gamma}$. Evaluate (39) at $(\gamma, \beta, y_0, y_1, y_2)$ and $(\tilde{\gamma}, \beta, y_0, y_1, y_2)$ and take difference:

$$\begin{aligned}&(y_1 - \tilde{\gamma} - \beta y_0)\triangle_{\tilde{\gamma},\gamma} g_1^* - (\tilde{\gamma} - \gamma) g_1^*(\gamma, \beta, y_0) \\ &+ (y_2 - \tilde{\gamma} - \beta y_1)\triangle_{\tilde{\gamma},\gamma} g_2^* - (\tilde{\gamma} - \gamma) g_2^*(\gamma, \beta, y_0, y_1) = 0\end{aligned} \tag{40}$$

where $\triangle_{\tilde{\gamma},\gamma} g_1^* = g_1^*(\tilde{\gamma}, \beta, y_0) - g_1^*(\gamma, \beta, y_0)$ and define $\triangle_{\tilde{\gamma},\gamma} g_2^*$ similarly.

In (40), y_2 only appears in the third term. Also, (40) must hold almost surely for all $\gamma, \tilde{\gamma}, \beta, y_0, y_1, y_2 \in \mathcal{C}_0$ such that $\gamma \neq \tilde{\gamma}$. Then it must be that, almost surely:

$$\triangle_{\tilde{\gamma},\gamma} g_2^* = 0 \tag{41}$$

If not, there exists a subset of $\mathcal{C}_0^5$ with positive Lebesgue measure in which $\triangle_{\tilde{\gamma},\gamma} g_2^* \neq 0$, and one can increase the value of y_2 for each element in this set to violate (40) with positive measure.

Now (41) implies that g_2^* is almost surely a constant function over γ:

$$g_2^*(\gamma, \beta, y_0, y_1) = g_2^*(\beta, y_0, y_1).$$

Take any $\gamma, \beta, \tilde{\beta}, y_0, y_1, y_2 \in \mathcal{C}$ such that $\beta \neq \tilde{\beta}$. Evaluate (39) at $(\gamma, \beta, y_0, y_1, y_2)$ and $(\gamma, \tilde{\beta}, y_0, y_1, y_2)$

and take difference:

$$\begin{aligned}&(y_1-\gamma-\tilde{\beta}y_0)\triangle_{\tilde{\beta},\beta}g_1^*-(\tilde{\beta}-\beta)y_0g_1^*(\gamma,\beta,y_0)\\&+(y_2-\gamma-\tilde{\beta}y_1)\triangle_{\tilde{\beta},\beta}g_2^*-(\tilde{\beta}-\beta)y_1g_2^*(\gamma,\beta,y_0,y_1)=\tilde{\beta}-\beta\end{aligned}\tag{42}$$

where $\triangle_{\tilde{\beta},\beta}g_1^*=g_1^*(\gamma,\tilde{\beta},y_0)-g_1^*(\gamma,\beta,y_0)$ and define $\triangle_{\tilde{\beta},\beta}g_2^*$ similarly. In (42), y_2 only appears in the third term. Therefore, almost surely:

$$g_2^*(\beta,y_0,y_1)=g_2^*(y_0,y_1),$$

with which (40) simplifies to

$$(y_1-\tilde{\gamma}-\beta y_0)\triangle_{\tilde{\gamma},\gamma}g_1^*-(\tilde{\gamma}-\gamma)g_1^*(\gamma,\beta,y_0)-(\tilde{\gamma}-\gamma)g_2^*(y_0,y_1)=0.\tag{43}$$

Let $\hat{\gamma}\in\mathcal{C}$ such that $\hat{\gamma}-\tilde{\gamma}=\tilde{\gamma}-\gamma$. Evaluate (43) at $(\gamma,\tilde{\gamma},\beta,y_0,y_1)$ and $(\tilde{\gamma},\hat{\gamma},\beta,y_0,y_1)$ and take difference:

$$(y_1-\hat{\gamma}-\beta y_0)\left(\triangle_{\hat{\gamma},\tilde{\gamma}}g_1^*-\triangle_{\tilde{\gamma},\gamma}g_1^*\right)-(\hat{\gamma}-\tilde{\gamma})\triangle_{\tilde{\gamma},\gamma}g_1^*-(\tilde{\gamma}-\gamma)\triangle_{\tilde{\gamma},\gamma}g_1^*=0.\tag{44}$$

In (44), y_1 only appears in the first term. Therefore, almost surely:

$$\triangle_{\hat{\gamma},\tilde{\gamma}}g_1^*-\triangle_{\tilde{\gamma},\gamma}g_1^*=0,$$

with which (44) reduces to

$$(\hat{\gamma}-\tilde{\gamma})\triangle_{\tilde{\gamma},\gamma}g_1^*+(\tilde{\gamma}-\gamma)\triangle_{\tilde{\gamma},\gamma}g_1^*=0.$$

Since $\hat{\gamma}-\tilde{\gamma}=\tilde{\gamma}-\gamma\neq 0$, this implies:

$$\triangle_{\tilde{\gamma},\gamma}g_1^*=0,$$

which means g_1^* is almost surely a constant over γ, i.e.

$$g_1^*(\gamma,\beta,y_0)=g_1^*(\beta,y_0).$$

Then I apply similar argument to (42), which yields:

$$g_1^*(\beta,y_0)=g_1^*(y_0),$$

with which (39) simplifies to

$$f^*(y_0, y_1, y_2) + g_1^*(y_0)(y_1 - \gamma - \beta y_0) + g_2^*(y_0, y_1)(y_2 - \gamma - \beta y_1) = \beta$$

almost surely for all $(\gamma, \beta, y_0, y_1, y_2)$. This is a linear identity in (γ, β), and so their coefficients must coincide. In other words, it must be that:

$$\begin{aligned} g_1^* + g_2^* &= 0, \\ y_0 g_1^* + y_1 g_2^* &= 1. \end{aligned}$$

Solving this for (g_1^*, g_2^*) yields:

$$g_1^* = \frac{1}{y_0 - y_1}, \quad g_2^* = \frac{1}{y_1 - y_0}.$$

However, g_1^* cannot be a function of y_1, which is contradiction. □

A.2 Proof of Theorem 1

This is a special case of Theorem 2. As discussed in Section 6.1.1, it suffices to consider $\lambda > 0$ for the upper bound and $\lambda < 0$ for the lower bound. □

A.3 Proof of Proposition 2

For the proof, it suffices to show that $[\tilde{L}, \tilde{U}]$ is the sharp identified set of μ_e under the assumptions. Then the inclusion $[L, U] \subseteq [\tilde{L}, \tilde{U}]$ follows from inclusion of the assumptions.

In what follows, I show that $\tilde{U}$ equals to the expression in the proposition. Similar argument applies to $\tilde{L}$.

By Theorem 2, the sharp upper bound $\tilde{U}$ is:

$$\tilde{U} = \min_{\lambda,\, \mu} \mathbb{E}\left(\max_{v} \left[e'v + \mu' \sum_{t=1}^{T} R_{it}(Y_{it} - R_{it}'v) + \lambda \sum_{t=1}^{T} (R_{it}'v)(Y_{it} - R_{it}'v) \right] \right)$$

where μ has the same dimension as R_{it}, and λ is scalar. To simplify notation, I abuse notation and let $\mathcal{R}_i = \sum_{t=1}^{T} R_{it} R_{it}'$ and $\mathcal{Y}_i = \sum_{t=1}^{T} R_{it} Y_{it}$. The $1/T$ scaling will be applied at the end of the proof.

With the notation, I can write $\tilde{U}$ concisely as

$$\tilde{U} = \min_{\mu,\lambda} \mathbb{E}\left(\max_{v}\left[e'v + \mu'\mathcal{Y}_i - \mu'\mathcal{R}_i v + \lambda\mathcal{Y}_i' v - v'\mathcal{R}_i v\right]\right).$$

The objective function of the inner maximization problem is a quadratic polynomial in v. As discussed in Section 6.1.1, it suffices to consider $\lambda > 0$, in which case $\tilde{U}$ simplifies because the inner maximization problem has a closed-form solution:

$$\tilde{U} = \min_{\lambda>0,\ \mu} \mathbb{E}\left(\mu'\mathcal{Y}_i + \frac{1}{4\lambda}\left[e + \lambda\mathcal{Y}_i - \mathcal{R}_i\mu\right]'\mathcal{R}_i^{-1}\left[e + \lambda\mathcal{Y}_i - \mathcal{R}_i\mu\right]\right).$$

I expand the terms:

$$\begin{aligned}\tilde{U} = \min_{\lambda>0,\ \mu}\Big[&\mu'\mathbb{E}(\mathcal{Y}_i) + \frac{1}{4\lambda}e'\mathbb{E}(\mathcal{R}_i^{-1})e + \frac{\lambda}{4}\mathbb{E}(\mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i) + \frac{1}{4\lambda}\mu'\mathbb{E}(\mathcal{R}_i)\mu \\ &+ \frac{1}{2}e'\mathbb{E}(\mathcal{R}_i^{-1}\mathcal{Y}_i) - \frac{1}{2\lambda}e'\mu - \frac{1}{2}\mu'\mathbb{E}(\mathcal{Y}_i)\Big].\end{aligned} \tag{45}$$

I solve for optimal μ given λ. The first order condition with respect to μ is

$$\mathbb{E}(\mathcal{Y}_i) + \frac{1}{2\lambda}\mathbb{E}(\mathcal{R}_i)\mu - \frac{1}{2\lambda}e - \frac{1}{2}\mathbb{E}(\mathcal{Y}_i) = 0.$$

Then the optimal μ that satisfies the first order condition is

$$\mu = \mathbb{E}(\mathcal{R}_i)^{-1}[e - \lambda\mathbb{E}(\mathcal{Y}_i)],$$

which I substitute into (45):

$$\begin{aligned}\tilde{U} = \min_{\lambda}\Big\{&[e - \lambda\mathbb{E}(\mathcal{Y}_i)]'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i) + \frac{1}{4\lambda}e'\mathbb{E}(\mathcal{R}_i^{-1})e + \frac{\lambda}{4}\mathbb{E}(\mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i) \\ &+ \frac{1}{4}\left[\frac{1}{\lambda}e'\mathbb{E}(\mathcal{R}_i)^{-1}e - 2e'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i) + \lambda\mathbb{E}(\mathcal{Y}_i)'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i)\right] \\ &+ \frac{1}{2}e'\mathbb{E}(\mathcal{R}_i^{-1}\mathcal{Y}_i) - \frac{1}{2}e'\mathbb{E}(\mathcal{R}_i)^{-1}\left[\frac{1}{\lambda}e - \mathbb{E}(\mathcal{Y}_i)\right] - \frac{1}{2}[e - \lambda\mathbb{E}(\mathcal{Y}_i)]'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i)\Big\}.\end{aligned}$$

The first order condition with respect to λ is

$$\frac{1}{\lambda^2}\left[e'\mathbb{E}(\mathcal{R}_i)^{-1}e - e'\mathbb{E}(\mathcal{R}_i^{-1})e\right] = \mathbb{E}(\mathcal{Y}_i)'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i) - \mathbb{E}(\mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i).$$

Since $\lambda > 0$, the optimal λ is:

$$\lambda = \sqrt{\frac{e'\mathbb{E}(\mathcal{R}_i^{-1})e - e'\mathbb{E}(\mathcal{R}_i)^{-1}e}{\mathbb{E}(\mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i) - \mathbb{E}(\mathcal{Y}_i)'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i)}}. \tag{46}$$

The numerator and the denominator inside the square root are both weakly positive, and they are zero if and only if $\mathcal{R}_i$ and $\mathcal{R}_i^{-1}\mathcal{Y}_i$ are non-degenerate across individuals, respectively. To see why, let R and Y be matrices that have the same dimensions as $\mathcal{R}_i$ and $\mathcal{Y}_i$, and define the functions:

$$\mathcal{E}(R) = e'R^{-1}e \quad \text{and} \quad \mathcal{D}(Y, R) = Y'R^{-1}Y.$$

One can show that $\mathcal{E}$ and $\mathcal{D}$ are convex functions. In particular, the following result is known[10]:

Lemma 2 (Kiefer, 1959, Lemma 3.2)**.** *For an integer $l > 0$, let $A_1, \ldots, A_l$ be $n \times m$ matrices and $B_1, \ldots, B_l$ be nonsingular positive definite and symmetric $n \times n$ matrices. Let $a_1, \ldots, a_l$ be positive real numbers such that $\sum_k a_k = 1$. Then*

$$\sum_{k=1}^{l} a_k A_k' B_k^{-1} A_k - \left[\sum_{k=1}^{l} a_k A_k\right]' \left[\sum_{k=1}^{l} a_k B_k\right]^{-1} \left[\sum_{k=1}^{l} a_k A_k\right] \geq 0$$

where '$\geq$' is the partial ordering defined in terms of positive semidefinite and positive definite matrices. In addition, the equality holds if and only if

$$B_1^{-1}A_1 = \ldots = B_l^{-1}A_l.$$

Applying Jensen's inequality to $\mathcal{E}(R)$ and $\mathcal{D}(Y, R)$ yields that the numerator and the denominator in (46) are weakly positive.

This finishes derivation of optimal μ and λ of (45). Plugging them in (45) and scaling the terms by $1/T$ gives the expression for $\tilde{U}$ in Proposition 2. □

A.4 Proof of Theorem 2

The proof focuses on showing (10). The same argument applies to (11).

In short, I show that (10) is the dual problem of (9). The main part of the proof is to formulate the problem into a standard form of infinite-dimensional linear programming and then show that the conditions for the duality theorem hold.

[10] See Nordström (2011) for its extension to complex field and generalized inverse.

I assume the following regularity conditions.

Assumption 5. The following conditions hold.

- $\mathcal{W} \times \mathcal{V}$ is compact.
- $(m, \phi_1, \ldots, \phi_K)$ are bounded Borel measurable functions on $\mathcal{W} \times \mathcal{V}$.
- There exists $P \in \mathcal{M}_{W \times V}$ that is a feasible point of (9) such that $\int m dP$ is finite and P has full support on $\mathcal{W} \times \mathcal{V}$.

The first condition is imposed for simplicity of technical argument. The second condition is mild given the the first condition. The third condition means that the data generating process of the model, or its observationally equivalent one, has full support. The third condition implies that (9) is feasible and that its solution is finite.

In what follows, I rewrite (9) into a standard form of infinite-dimensional linear programming, for which I introduce additional notation. Recall that $\mathcal{M}_{W \times V}$ is a linear space of bounded and finitely additive signed Borel measures on $\mathcal{W} \times \mathcal{V}$. Let $\overline{\mathcal{F}}_{W \times V}$ be the dual space of $\mathcal{M}_{W \times V}$, and let $\mathcal{F}_{W \times V}$ be the space of all bounded Borel measurable functions on $\mathcal{W} \times \mathcal{V}$. Then $\mathcal{F}_{W \times V}$ is a linear subspace of $\overline{\mathcal{F}}_{W \times V}$ because it is the double dual of $\mathcal{F}_{W \times V}$.

For $P \in \mathcal{M}_{W \times V}$ and $f \in \overline{\mathcal{F}}_{W \times V}$, define the *dual pairing*

$$\langle P, f \rangle = \int f dP.$$

Let $\mathcal{M}_W$ be the projection of $\mathcal{M}_{W \times V}$ onto $\mathcal{W}$. Let $\overline{\mathcal{F}}_W$ be the dual space of $\mathcal{M}_W$, and let $\mathcal{F}_W$ be the space of all bounded Borel measurable functions on $\mathcal{W}$. Then $\mathcal{F}_W$ is a linear subspace of $\overline{\mathcal{F}}_W$. In addition, define $\mathcal{G} = \mathbb{R}^K \times \mathcal{M}_W$ and $\mathcal{H} = \mathbb{R}^K \times \overline{\mathcal{F}}_W$, and let $g = (g_1, \ldots, g_K, P_g)$ and $h = (h_1, \ldots, h_K, f_h)$ be their generic elements. Note that $\mathcal{H}$ is the dual space of $\mathcal{G}$. Define the dual pairing

$$\langle g, h \rangle = \sum_{k=1}^{K} g_k h_k + \int f_h dP_g.$$

Define a linear map $A : \mathcal{M}_{W \times V} \mapsto \mathcal{G}$ where

$$A(P) = \left(\int \phi_1 dP, \ldots, \int \phi_K dP, P(\cdot, \mathcal{V}) \right).$$

Then

$$\langle A(P), h \rangle = \sum_{k=1}^{K} h_k \int \phi_k dP + \int_{\mathcal{W}} f_h(w) P(dw, \mathcal{V}).$$

It is straightforward to show:

$$\int_{\mathcal{W}} f_h(w) P(dw, \mathcal{V}) = \int_{\mathcal{W}\times\mathcal{V}} f_h(w) dP(w, v).$$

Then:

$$\langle A(P), h\rangle = \sum_{k=1}^{K} h_k \int \phi_k dP + \int f_h dP = \int \left[\sum_{k=1}^{K} h_k \phi_k + f_h\right] dP \equiv \langle P, A^*(h)\rangle, \tag{47}$$

where $A^*(h) : \mathcal{H} \mapsto \overline{\mathcal{F}}_{W\times V}$ is defined as

$$A^*(h) = \sum_{k=1}^{K} h_k \phi_k + f_h.$$

Equation (47) shows that A is weakly continuous and that A^* is the adjoint of A. This is a key condition for the duality theorem.

With these notation, I rewrite (9) into a standard form of infinite-dimensional linear programming:

$$\min_{P\in\mathcal{M}_{W\times V}} \langle P, m\rangle \qquad \text{subject to} \qquad A(P) = c, \qquad P \geq 0, \tag{48}$$

where $c = (0, \ldots, 0, P_W)$.

With Assumption 5 and A being weakly continuous, the strong duality holds[11]. That is, the optimal solution to (48) equals to the solution to:

$$\max_{h\in\mathcal{H}} \langle c, h\rangle \qquad \text{subject to} \qquad m - A^*(h) \geq 0, \qquad P \geq 0,$$

which I can write more concretely as:

$$\max_{h_1,\ldots,h_K\in\mathbb{R},\ f_h\in\overline{\mathcal{F}}_W} \int f_h dP_W \qquad \text{subject to} \qquad \sum_{k=1}^{K} h_k \phi_k + f_h \leq m. \tag{49}$$

Now I simplify (49) and obtain (10). First, rearrange the constraint of (49):

$$f_h(w) \leq m(w, v) - \sum_{k=1}^{K} h_k \phi_k(w, v).$$

[11] See e.g. Anderson (1983) and an appendix chapter of Lasserre (2010).

The left-hand side is only a function of w. Therefore:

$$f_h(w) \leq \min_{v \in \mathcal{V}} \left[m(w,v) - \sum_{k=1}^{K} h_k \phi_k(w,v) \right] \quad \text{for all} \quad w \in \mathcal{W}.$$

The objective of (49) is to maximize the integral of $f_h(w)$. Therefore, f_h^* must satisfy:

$$f_h^*(w) = \min_{v \in \mathcal{V}} \left[m(w,v) - \sum_{k=1}^{K} h_k \phi_k(w,v) \right] \tag{50}$$

almost surely on $\mathcal{W}$. If not, i.e. if

$$f_h^*(w) < \min_{v \in \mathcal{V}} \left[m(w,v) - \sum_{k=1}^{K} h_k \phi_k(w,v) \right]$$

with positive probability, then I can increase the value of f_h^* by an infinitesimal amount which increases the value of the objective, which is contradiction.

Now I substitute (50) into (49):

$$\max_{h_1,\ldots,h_K \in \mathbb{R}} \int \min_{v \in \mathcal{V}} \left[m(w,v) - \sum_{k=1}^{K} h_k \phi_k(w,v) \right] dP_W(w).$$

The above display remains equivalent even if I switch the signs of $(h_1, \ldots, h_K)$ because the h's are choice variables supported on $\mathbb{R}^K$. If I do so, the problem becomes:

$$\max_{h_1,\ldots,h_K \in \mathbb{R}} \int \min_{v \in \mathcal{V}} \left[m(w,v) + \sum_{k=1}^{K} h_k \phi_k(w,v) \right] dP_W(w)$$

which is the expression in (10). □

A.5 Proof of Proposition 3

By the proof of Proposition 6, $\hat{L}(\lambda)$ and $L(\lambda)$ are concave. Then $\hat{L}$ uniformly converges to L on any compact set $K \subseteq \mathbb{R}^K$ (as in the proof of Theorem 2.7 in Newey and McFadden (1994)):

$$\sup_{\lambda \in K} |\hat{L}(\lambda) - L(\lambda)| \xrightarrow{p} 0. \tag{51}$$

Let $\hat{\lambda} = \text{argmax}_\lambda \, \hat{L}(\lambda)$ and $\lambda_0 = \text{argmax}_\lambda \, L(\lambda)$. If there are multiple argmax's, choose any

of them. Then:

$$\hat{L}(\hat{\lambda}) \geq \hat{L}(\lambda_0).$$

For $\hat{\lambda}$ that is on a compact set $K \subseteq \mathbb{R}^K$:

$$\begin{aligned}
|L(\lambda_0) - \hat{L}(\hat{\lambda})| &\leq L(\lambda_0) - L(\hat{\lambda}) + |L(\hat{\lambda}) - \hat{L}(\hat{\lambda})| \\
&= \hat{L}(\lambda_0) - L(\hat{\lambda}) + |L(\hat{\lambda}) - \hat{L}(\hat{\lambda})| + o_p(1) \\
&\leq \hat{L}(\hat{\lambda}) - L(\hat{\lambda}) + |L(\hat{\lambda}) - \hat{L}(\hat{\lambda})| + o_p(1) \\
&\leq 2|L(\hat{\lambda}) - \hat{L}(\hat{\lambda})| + o_p(1) = o_p(1),
\end{aligned}$$

where the last equality follows from (51).

Let Λ_0 be the set of all $\text{argmax}_\lambda L(\lambda)$. Let K_0 be a compact set containing an open neighborhood of Λ_0 with radius $\varepsilon > 0$. If such ε does not exist, it means that $L(\lambda)$ is a constant function, in which case consistency is immediate. If such ε exists, by Theorem 5.14 of Van der Vaart (2000):

$$\mathbb{P}(\tilde{d}(\hat{\lambda}, \Lambda_0) \geq \varepsilon \wedge \hat{\lambda} \in K_0) \longrightarrow 0$$

where $\tilde{d}(\hat{\lambda}, \Lambda_0) = \inf\{d(\hat{\lambda}, \lambda) \mid \lambda \in \Lambda_0\}$ and d is Euclidean distance. This implies $\hat{\lambda} \in K_0$ with probability approaching to one. □

A.6 Proof of Proposition 4

I can rewrite (21) as:

$$\min_{P \in \mathcal{M}_{W \times V},\, P \geq 0,\, \delta \geq 0} \delta \qquad \text{subject to} \qquad \begin{aligned}
&\int dP = 1, \\
&\int \phi_k(W_i, V_i) dP \leq \delta, \quad k = 1, \ldots, K, \\
&\int \phi_k(W_i, V_i) dP \geq -\delta, \quad k = 1, \ldots, K, \\
&\int P(w, dV_i) = \hat{P}_W(w) \ \text{ for all } w \in \mathcal{W}.
\end{aligned}$$

Then I can replicate the argument of Theorem 2 and show that (22) follows by taking the dual and simplifying it. □

A.7 Proof of Proposition 5

I can rewrite (24) as:

$$\min_{P\in\mathcal{M}_{W\times V},\ P\geq 0}\int m(W_i,V_i)dP \qquad \text{subject to} \qquad \begin{aligned}&\int \phi_k(W_i,V_i)dP\leq \delta^*, \quad k=1,\dots,K,\\ &\int \phi_k(W_i,V_i)dP\geq -\delta^*, \quad k=1,\dots,K,\\ &\int P(w,dv)=\hat{P}_W(w) \ \text{ for all } w\in\mathcal{W}.\end{aligned}$$

Then I can replicate the argument of Theorem 2 and show that (23) follows by taking the dual and simplifying it. □

A.8 Proof of Proposition 6

It suffices to show that G_L is concave in λ. Let $w \in \mathcal{W}$, and let $\lambda_1 = (\lambda_{11},\dots,\lambda_{1K})$ and $\lambda_2 = (\lambda_{21},\dots,\lambda_{2K})$ be two distinct points in $\mathbb{R}^K$. Then, for any $t\in[0,1]$:

$$\begin{aligned}&G_L(t\lambda_1+(1-t)\lambda_2,w)\\ &=\min_{v\in\mathcal{V}}\left\{t\left[m(w,v)+\sum_{k=1}^K\lambda_{1k}\phi_k(w,v)\right]+(1-t)\left[m(w,v)+\sum_{k=1}^K\lambda_{2k}\phi_k(w,v)\right]\right\}\\ &\geq t\min_{v\in\mathcal{V}}\left\{m(w,v)+\sum_{k=1}^K\lambda_{1k}\phi_k(w,v)\right\}+(1-t)\min_{v\in\mathcal{V}}\left\{m(w,v)+\sum_{k=1}^K\lambda_{2k}\phi_k(w,v)\right\}\\ &=tG_L(\lambda_1,w)+(1-t)G_L(\lambda_2,w).\end{aligned}$$

This is the definition of concavity. □

B Identification under conditional moment restrictions

This section extends the model of Assumption 4 and consider conditional moment restrictions. Consider the following model:

Assumption 6. The random vectors (W_i,V_i) satisfy

$$\begin{aligned}\mathbb{E}(\phi_k(W_i,V_i))&=0, \quad k=1,\dots,K_U,\\ \mathbb{E}(\psi_k(W_i,V_i)|A_{ik})&=0, \quad k=1,\dots,K_C,\end{aligned}$$

where ϕ_k's and ψ_k's are real-valued moment functions, $A_{i1}, \ldots, A_{i,K_C}$ are subvectors of (W_i, V_i) and K_U and K_C are the number of unconditional and conditional moment restrictions, respectively.

This subsection obtains the counterpart of Theorem 2 under Assumption 6. In other words, under Assumption 6, I characterize the identified set of

$$\theta = \mathbb{E}(m(W_i, V_i))$$

for some known function $m : \mathcal{W} \times \mathcal{V} \mapsto \mathbb{R}$.

For each A_{ik}, which is a subvector of (W_i, V_i), let A'_{ik} be the vector that collects the remaining variables of (W_i, V_i). I abuse notation and write any function $f(w, v)$ on $\mathcal{W} \times \mathcal{V}$ equivalently as $f(a_k, a'_k)$ on $\mathcal{A}_k \times \mathcal{A}'_k$, where $\mathcal{A}_k$ is the support of A_{ik}, and let $\mathcal{A}'_k$ is the support of A'_{ik}.

The following characterizes the identified set I of θ.

Theorem 3. *Suppose Assumption 6 hold. Suppose also that (W_i, V_i) follows a σ-finite distribution that is absolutely continuous with respect to the Lebesgue measure. Then, under suitable additional regularity conditions, $I = [L, U]$ where, for $\lambda_k \in \mathbb{R}$ for $k = 1, \ldots, K_U$ and $\mu_k : \mathcal{A}_k \mapsto \mathbb{R}$ for $k = 1, \ldots, K_C$,*

$$L = \max_{\{\lambda_k\}_{k=1}^{K_U}, \{\mu_k\}_{k=1}^{K_C}} \mathbb{E}\left[\min_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K_U} \lambda_k \phi_k(W_i, v) + \sum_{k=1}^{K_C} \mu_k(A_k(W_i, v))\psi_k(W_i, v) \right\}\right] \quad (52)$$

and

$$U = \min_{\{\lambda_k\}_{k=1}^{K_U}, \{\mu_k\}_{k=1}^{K_C}} \mathbb{E}\left[\max_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K_U} \lambda_k \phi_k(W_i, v) + \sum_{k=1}^{K_C} \mu_k(A_k(W_i, v))\psi_k(W_i, v) \right\}\right] \quad (53)$$

where $A_k(w, v)$ is the value of A_{ik} given $W_i = w$ and $V_i = v$.

Proof. The proof focuses on showing (52). The same argument applies to (53).

Using absolute continuity, I identify an element of $\mathcal{M}_{W \times V}$ by its density $p : \mathcal{W} \times \mathcal{V} \mapsto \mathbb{R}$.

Let p_W be the density of P_W. The identified set I is defined by

$$
\begin{aligned}
I \equiv \Bigg\{ \int m(w,v)p(w,v)\,d(w,v) \;\Big|\; & p \in \mathcal{M}_{W\times V}, \quad p \geq 0, \\
& \int \phi_k(w,v)p(w,v)d(w,v) = 0, \quad k = 1,\ldots,K_U, \\
& \int \psi_k(a_k,a_k')p(a_k,a_k')da_k' = 0 \text{ for all } a_k \in \mathcal{A}_k, \quad k = 1,\ldots,K_C, \\
& \int p(w,v)dv = p_W(w) \;\; \text{for all } w \in \mathcal{W} \Bigg\},
\end{aligned}
$$

where a_k is an element of $\mathcal{A}_k$ and a_k' is an element of $\mathcal{A}_k'$. Note that the second line represents unconditional moment restrictions and that the third line represents conditional moment restrictions.

The lower bound of I is given by the infinite-dimensional linear program

$$
\begin{aligned}
\min_{p\in\mathcal{M}_{W\times V},\, p\geq 0} & \int m(w,v)p(w,v)d(w,v) \qquad \text{subject to} \\
& \int \phi_k(w,v)p(w,v)d(w,v) = 0, \quad k = 1,\ldots,K_U, \\
& \int \psi_k(a_k,a_k')p(a_k,a_k')da_k' = 0, \text{ for all } a_k \in \mathcal{A}_k, \quad k = 1,\ldots,K_C, \\
& \int p(w,v)dv = p_W(w) \;\; \text{for all } w \in \mathcal{W}.
\end{aligned}
\tag{54}
$$

In what follows, I show that (52) is the dual problem of (54). I assume the following regularity conditions:

Assumption 7. The following conditions hold.

- $\mathcal{W} \times \mathcal{V}$ is compact.
- $(m, \phi_1, \ldots, \phi_{K_U}, \psi_1, \ldots, \psi_{K_C})$ are L^∞ with respect to the Lebesgue measure.
- There exists $p \in \mathcal{M}_{W\times V}$ that is a feasible point of (54) such that $\int m(w,v)p(w,v)d(w,v)$ is finite and $p > 0$ on $\mathcal{W} \times \mathcal{V}$.
- Every density function $p \in \mathcal{M}_{W\times V}$ is L^∞ with respect to the Lebesgue measure.

The first three conditions of Assumption 7 are the counterparts of Assumption 5. Note that, since $\mathcal{W} \times \mathcal{V}$ is compact, the second condition implies they are L^p for any $p \geq 1$. The fourth condition is restrictive, but it is useful enough for showing non-identification results of Proposition 1.

To show the proof, I introduce additional notation. Let $L^2(\mathcal{W} \times \mathcal{V})$ be the space of all L^2 functions on $\mathcal{W} \times \mathcal{V}$, and let $L^2(\mathcal{W})$ be the space of all L^2 functions on $\mathcal{W}$. We also let $L^2(\mathcal{A}_k)$ be the space of all L^2 functions on $\mathcal{A}_k$.

Define $\mathcal{G} = \mathcal{H} = \mathbb{R}^K \times L^2(\mathcal{A}_1) \times \ldots \times L^2(\mathcal{A}_{K_C}) \times L^2(\mathcal{W})$. Denote their generic elements as $g = (g_1, \ldots, g_{K_U}, \overline{g}_1, \ldots, \overline{g}_{K_C}, f_g)$ and $h = (h_1, \ldots, h_{K_U}, \overline{h}_1, \ldots, \overline{h}_{K_C}, f_h)$, respectively. Note that $\mathcal{H}$ is a dual space of $\mathcal{G}$.

Define the linear map $A : \mathcal{M}_{W \times V} \mapsto \mathcal{G}$:

$$A(p) = \left(\int \phi_1 p \; d(w,v), \ldots, \int \phi_K p \; d(w,v), \int \psi_k p \; da'_1, \ldots, \int \psi_k p \; da'_{K_C}, \int p \; dv \right).$$

Define the dual pairing:

$$\langle A(P), h \rangle = \sum_{k=1}^{K_U} h_k \int \phi_k p \; d(w,v) + \sum_{k=1}^{K_C} \int \int \psi_k p \; da'_k \, \overline{h}_k da_k + \int f_h \int p dv dw.$$

It is straightforward to show:

$$\iint \psi_k p \; da'_k \, \overline{h}_k da_k = \int \psi_k \overline{h}_k p \; d(w,v)$$

and

$$\int f_h \int p dv dw = \int f_h p \; d(w,v).$$

Then:

$$\langle A(P), h \rangle = \int \left[\sum_{k=1}^{K_U} h_k \phi_k + \sum_{k=1}^{K_C} \overline{h}_k \psi_k + f_h \right] p(w,v) d(w,v). \equiv \langle p, A^*(h) \rangle, \tag{55}$$

where $A^*(h) : \mathcal{H} \mapsto L^2(\mathcal{W} \times \mathcal{V})$ is defined as

$$A^*(h) = \sum_{k=1}^{K_U} h_k \phi_k + \sum_{k=1}^{K_C} \overline{h}_k \psi_k + f_h.$$

Equation (55) shows that A is weakly continuous and that A^* is the adjoint of A. Note that $\overline{h}_k \psi_k$ is L^2 since ψ_k is L^∞ and hence bounded almost surely.

Then, similarly to the proof of Theorem 2, the strong duality holds under Assumption 7 and the weak continuity of A. The optimal solution to (54) equals to the solution to:

$$\max_{h_1, \ldots, h_{K_U}, \overline{h}_1, \ldots, \overline{h}_{K_C}, f_h} \int f_h(w) p_w(w) dw \qquad \text{subject to} \qquad \sum_{k=1}^{K_U} h_k \phi_k + \sum_{k=1}^{K_C} \overline{h}_k \psi_k + f_h \leq m. \tag{56}$$

Similarly to the proof of Theorem 2, the optimal solution f_h^* must satisfy:

$$f_h(w) = \min_{v \in \mathcal{V}} \left[m(w,v) - \sum_{k=1}^{K_U} h_k \phi_k(w,v) - \sum_{k=1}^{K_C} \overline{h}_k(A_k(w,v)) \psi_k(w,v) \right].$$

Switching the signs of $h_1, \ldots, h_{K_U}, \overline{h}_1, \ldots, \overline{h}_{K_C}$ yields the expression in (52).

□

C Global polynomial optimization

This section formally discusses the theory of global polynomial optimization introduced in Section 6.1.2. The theory can be extended to global optimization of semi-algebraic functions, which include polynomials and the functions created by their addition, subtraction, multiplication, division, $\max\{\cdot,\cdot\}$, $\min\{\cdot,\cdot\}$, absolute value, square root, cubic root, etc.

C.1 Setup and notation

Let $v = (v_1, \ldots, v_m) \in \mathbb{R}^m$. Let $\mathcal{V}$ be the vector space of polynomials in v over the field of real numbers. The canonical basis of $\mathcal{V}$ is the set of all monomials in v:

$$\{v_1^{\alpha_1} v_2^{\alpha_2} \cdots v_m^{\alpha_m} \mid \alpha_1, \ldots, \alpha_m \in \mathbb{N}\} \quad \text{where} \quad \mathbb{N} = \{0, 1, 2, \ldots\}.$$

Let $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{N}^m$ and write monomials concisely as $v^\alpha = v_1^{\alpha_1} v_2^{\alpha_2} \cdots v_m^{\alpha_m}$. Let $|\alpha| = \alpha_1 + \cdots + \alpha_m$ be the degree of v^α.

The standard form of a polynomial $f(v)$ is:

$$f(v) = \sum_{\alpha \geq 0} c_\alpha v^\alpha$$

where $\{c_\alpha\}_{\alpha \in \mathbb{N}^m}$ is a sequence of real numbers indexed by α.

Let f be a polynomial in v whose degree is $d < \infty$. I am interested in minimizing f with respect to v:

$$\min_{v \in \mathbb{K}} f(v) = \min_{v \in \mathbb{K}} \sum_{\alpha \,:\, |\alpha| \leq d} c_\alpha v^\alpha. \tag{57}$$

The domain $\mathbb{K}$ is either $\mathbb{K} = \mathbb{R}^m$ or

$$\mathbb{K} = \{v \in \mathbb{R}^m \mid g_j(v) \geq 0,\ j = 1, \ldots, J\} \tag{58}$$

where g_js are polynomials in v. If $\mathbb{K}$ is given in the form of (58), assume that $\mathbb{K}$ is compact. Let $d_j < \infty$ be the degree of g_j.

Example 8. Let $J = 1$. Let $g_1 = 1 - \sum_{k=1}^{m} v_k^2$. Then $\mathbb{K}$ is a unit sphere in $\mathbb{R}^m$.

Example 9. Let $J = 2m$. Let $g_k = 1 - v_k$ for $k = 1, \ldots, m$ and $g_{m+k} = v_k$ for $k = 1, \ldots, m$. Then $\mathbb{K}$ is a unit rectangle $[0, 1]^m$ in $\mathbb{R}^m$.

The semidefinite programming (SDP) approach of polynomial optimization transforms (57) into a convex optimization problem[12]. Let $\tilde{\mathcal{V}}$ be the space of Borel measures in $\mathbb{R}^m$ whose supports are contained in $\mathbb{K}$. I reformulate (57) as:

$$\min_{v \in \mathbb{K}} f(v) = \min_{P \in \tilde{\mathcal{V}},\ \int dP = 1} \int f(v) dP = \min_{P \in \tilde{\mathcal{V}},\ \int dP = 1} \sum_{\alpha\ :\ |\alpha| \leq d} c_\alpha \int v^\alpha dP. \tag{59}$$

The optimal P of (59) is the point-mass distribution concentrated at the minimizer of $f(v)$. Let $\{y_\alpha\}$ be a sequence of real numbers indexed by α. I rewrite (59) as:

$$\min_{\{y_\alpha\}} \sum_{\alpha} c_\alpha y_\alpha \qquad \text{subject to} \qquad y_\alpha = \int v^\alpha dP \quad \text{for some } P \in \tilde{\mathcal{V}},\ \int dP = 1. \tag{60}$$

The objective function of (60) linear in $\{y_\alpha\}$, which means it is a convex function of $\{y_\alpha\}$. If I also characterize the constraint of (60) as a convex constraint in $\{y_\alpha\}$, then (60) becomes a convex optimization problem. The next subsection discusses how I characterize the constraint.

C.2 The moment problem

The constraint of (60) is closely related to the *moment problem* in mathematics. The moment problem asks the following question: "Given the infinite real number sequence $\{y_\alpha\}_{\alpha \in \mathbb{N}^m}$, does there exist a measure P supported on $\mathbb{R}^m$ such that $y_\alpha = \int v^\alpha dP$ for all $\alpha \geq 0$?" In words, it asks whether a sequence of numbers can be justified as moments of some distribution. If the answer is yes, we say that $\{y_\alpha\}_{\alpha \in \mathbb{N}^m}$ has a *representing measure*.

Given an infinite sequence $\mathbf{y} = \{y_\alpha\}_{\alpha \in \mathbb{N}^m}$, define the linear functional $L_{\mathbf{y}} : \mathcal{V} \mapsto \mathbb{R}$ by:

$$f = \sum_{\alpha} c_\alpha v^\alpha \quad \mapsto \quad L_{\mathbf{y}}(f) = \sum_{\alpha} c_\alpha y_\alpha.$$

The following is a key result for the moment problem.

[12] See Lasserre (2010) and Lasserre (2015) for reference.

Theorem 4 (Riesz-Haviland). *Let* $\mathbf{y} = \{y_\alpha\}_{\alpha \in \mathbb{N}^m}$ *and let* $\mathbb{K} \subseteq \mathbb{R}^m$ *be closed. There exists a finite Borel measure* $P \in \tilde{\mathcal{V}}$ *such that*

$$y_\alpha = \int v^\alpha dP \quad \text{for all } \alpha \in \mathbb{N}^m$$

if and only if $L_{\mathbf{y}}(f) \geq 0$ *for all polynomials* $f \in \mathcal{V}$ *nonnegative on* $\mathbb{K}$.

Riesz-Haviland theorem characterizes the constraint of (60) in terms of $\mathbf{y} = \{y_\alpha\}_{\alpha \in \mathbb{N}^m}$. However, the characterization is not useful yet because checking $L_{\mathbf{y}}(f) \geq 0$ for all nonnegative f is computationally infeasible. In what follows, I derive a tractable characterization from the theorem.

Let $\mathbb{N}^m_r = \{\alpha \in \mathbb{N}^m,\ |\alpha| \leq r\}$ and let $\mathcal{V}_r$ be the space of polynomials in v whose degree is at most r. The canonical basis of $\mathcal{V}_r$ is:

$$u_r(v) = (1, v_1, \ldots, v_m, v_1^2, v_1 v_2, \ldots, v_m^2, \ldots, v_1^r, \ldots, v_m^r)'.$$

Let s_r be the length of $u_r(v)$, which equals $(n+r)!/n!r!$ ($n+r$ choose n). Note that each element of $\mathcal{V}_r$, a polynomial whose degree is at most r, is identified by its coefficient vector of length s_r.

Let $\mathbf{y} = \{y_\alpha\}_{\alpha \in \mathbb{N}^m}$. The *moment matrix* of dimension s_r, denoted by $M_r(\mathbf{y})$, is defined by:

$$M_r(\mathbf{y}) = L_{\mathbf{y}}(u_r(v) u_r(v)')$$

where we apply $L_{\mathbf{y}}$ element-wise. Note that $u_r(v)u_r(v)'$ is a matrix of polynomials whose degrees are at most $2r$. Equivalently, the moment matrix is a square matrix labeled by $\alpha, \beta \in \mathbb{N}^m_d$ such that

$$[M_r(\mathbf{y})]_{\alpha,\beta} = L_y(v^\alpha v^\beta) = y_{\alpha+\beta}.$$

Example 10. If $m = r = 2$, then the moment matrix $M_2(\mathbf{y})$ is a 6×6 matrix given by

$$M_2(\mathbf{y}_4) = \left(\begin{array}{c|cc|ccc} y_{00} & y_{10} & y_{01} & y_{20} & y_{11} & y_{02} \\ \hline y_{10} & y_{20} & y_{11} & y_{30} & y_{21} & y_{12} \\ y_{01} & y_{11} & y_{02} & y_{21} & y_{12} & y_{03} \\ \hline y_{20} & y_{30} & y_{21} & y_{40} & y_{31} & y_{22} \\ y_{11} & y_{21} & y_{12} & y_{31} & y_{22} & y_{13} \\ y_{02} & y_{12} & y_{03} & y_{22} & y_{13} & y_{04} \end{array}\right).$$

where $y_{\alpha_1\alpha_2} = v_1^{\alpha_1} v_2^{\alpha_2}$.

Let $p, q \in \mathcal{V}_r$, and let their coefficient vectors be $\mathbf{p}$ and $\mathbf{q}$. The following result is known:

$$L_{\mathbf{y}}(pq) = \mathbf{p}' M_r(\mathbf{y}) \mathbf{q}.$$

Suppose $\mathbf{y}$ has a representing measure. Then:

$$\mathbf{p}' M_r(\mathbf{y}) \mathbf{p} = L_{\mathbf{y}}(p^2) \geq 0$$

by Riesz-Haviland theorem, which implies $M_r(\mathbf{y})$ is positive semidefinite (PSD). Therefore, $M_r(\mathbf{y})$ being PSD is a necessary condition for $\mathbf{y}$ having a representing measure.

For $g_j \in \mathcal{V}$ whose degree is $2r_j$ or $2r_j - 1$, the *localizing matrix* of dimension s_{r-r_j} with respect to g_j and $\mathbf{y}$, denoted by $M_{r-r_j}(g_j \mathbf{y})$, is defined by:

$$M_{r-r_j}(g_j \mathbf{y}) = L_{\mathbf{y}}(g_j(v) u_r(v) u_r(v)').$$

Equivalently, the localizing matrix is a square matrix labeled by $\alpha, \beta \in \mathbb{N}^m_{r_j}$ such that

$$[M_{r-r_j}(g_j \mathbf{y})]_{\alpha,\beta} = L_{\mathbf{y}}(g_j v^\alpha v^\beta).$$

Example 11. If $m = 2$, $r = 2$ and $g_1(v) = 1 - v_1^2 - v_2^2$, then the localizing matrix $M_1(g_1 \mathbf{y})$ is:

$$M_1(g_1 \mathbf{y}) = \begin{pmatrix} y_{00} & y_{10} & y_{01} \\ y_{10} & y_{20} & y_{11} \\ y_{01} & y_{11} & y_{02} \end{pmatrix} - \begin{pmatrix} y_{20} & y_{30} & y_{21} \\ y_{30} & y_{40} & y_{31} \\ y_{21} & y_{31} & y_{22} \end{pmatrix} - \begin{pmatrix} y_{02} & y_{12} & y_{03} \\ y_{12} & y_{22} & y_{13} \\ y_{03} & y_{13} & y_{04} \end{pmatrix}.$$

where $y_{\alpha_1 \alpha_2} = v_1^{\alpha_1} v_2^{\alpha_2}$.

Similar to the case of moment matrix, if $\mathbf{y}$ has a representing measure on $\mathbb{K}$ such that $g_j \geq 0$ on $\mathbb{K}$, then:

$$\mathbf{p}' M_{r-r_j}(g_j \mathbf{y}) \mathbf{p} = L_{\mathbf{y}}(g_j p^2) \geq 0$$

by Riesz-Haviland theorem. Therefore, $M_{r-r_j}(g_j \mathbf{y})$ being PSD is a necessary condition for $\mathbf{y}$ having a representing measure on $\mathbb{K}$ such that $g_j \geq 0$ on $\mathbb{K}$.

C.3 Constrained polynomial optimization

The necessary conditions from the previous subsection become equivalent conditions in the case that $\mathbb{K}$ has the form (58) and is compact. Note that, if $\mathbb{K} \in \mathbb{R}^m$ is compact, there exists a

real number $B > 0$ such that

$$B - ||v||^2 = B - \sum_{k=1}^{m} v_k^2 \geq 0 \quad \text{on } \mathbb{K}.$$

The following provides a necessary and sufficient condition for the moment problem.

Theorem 5 (Putinar's Positivstellensatz)**.** *Let $\mathbb{K}$ be defined as in (58) and suppose $\mathbb{K}$ is compact. Let $g_{J+1} = B - \sum_{k=1}^{m} v_k^2$. Then $\mathbf{y}$ has a finite Borel representing measure whose support is contained in $\mathbb{K}$ if and only if the following conditions hold:*

- *$M_r(\mathbf{y})$ is PSD for all $r \geq 1$,*
- *$M_r(g_j\mathbf{y})$ is PSD for $j = 1, \ldots, J+1$, for all $r \geq 1$.*

Putinar's Positivstellensatz implies that, when $\mathbb{K}$ has the form (58) and is compact, the polynomial optimization problem of (60) is equivalent to:

$$\begin{aligned} V = \min_{\mathbf{y}} \sum_{\alpha} c_\alpha y_\alpha \qquad \text{subject to} \qquad & y_0 = 1, \\ & M_r(\mathbf{y}) \text{ is PSD for all } r, \\ & M_r(g_j\mathbf{y}) \text{ is PSD for } j = 1, \ldots, J+1, \text{ for all } r. \end{aligned} \tag{61}$$

The PSD constraints are convex constraints because the set of semidefinite matrices is a convex set (see Lemma 2 in the proof of Proposition 2). Therefore, (61) is a convex optimization problem. Moreover, the constraints belong to a special class of convex constraints so that (61) has its own name: semidefinite program (SDP).

Yet, (61) has infinite number of constraints, and so it is not computationally feasible. To obtain a computationally feasible problem, let $r_j = \lfloor (d_j + 1)/2 \rfloor$ and choose $r \geq \lfloor (d+1)/2 \rfloor$. Let $\mathbf{y}_{2r} = \{y_\alpha\}_{\alpha \in \mathbb{N}^m_{2r}}$. Consider the problem

$$\begin{aligned} V_r = \min_{\mathbf{y}_{2r}} \sum_{|\alpha| \leq 2r} c_\alpha y_\alpha \qquad \text{subject to} \qquad & y_0 = 1, \\ & M_r(\mathbf{y}_{2r}) \text{ is PSD}, \\ & M_{r-r_j}(g_j\mathbf{y}_{2r}) \text{ is PSD for all } j = 1, \ldots, J+1. \end{aligned} \tag{62}$$

Note that $M_r(\mathbf{y}_{2r})$ being PSD implies $M_{r'}(\mathbf{y}_{2r})$ being PSD for all $r' \leq r$. (62) has finite number of constraints, and it can be solved numerically using softwares available in the industry.

The constraint of (62) is a finite subset of the constraint of (61). This means that $V_r \leq V$.

Also, V_r is monotonically increasing in r since there are more constraints as r increases. It is known that $V_r \nearrow V$ as $r \to \infty$ (Lasserre, 2010, Theorem 4.1).

The convergence is finite under suitable conditions (Lasserre, 2015, Theorem 6.5), that is, $V_r = V$ for some finite r. I discuss a procedure for checking $V_r = V$, which is called *certificate* of optimality.

The procedure asks the following question: "Given a finite sequence $\mathbf{y}_{2r} = \{y_\alpha\}_{\alpha \in \mathbb{N}^m_{2r}}$ such that the moment matrix $M_r(\mathbf{y}_{2r})$ is PSD, can we find new numbers $\{y_\alpha\}_{\alpha\,:\,2r<|\alpha|\leq 2r+2}$ such that $M_{r+1}(\mathbf{y}_{2r+2})$ is PSD?" If it is possible, then $M_{r+1}(\mathbf{y}_{2r+2})$ is called a *positive extension* of $M_r(\mathbf{y}_{2r})$. Moreover, if in addition $\operatorname{rank}(M_r(\mathbf{y}_{2r})) = \operatorname{rank}(M_{r+1}(\mathbf{y}_{2r+2}))$ holds, then $M_{r+1}(\mathbf{y}_{2r+2})$ is called a *flat extension* of $M_r(\mathbf{y}_{2r})$.

A measure is called *s-atomic* if it is a discrete measure with s support points. The following theorem holds:

Theorem 6 (Lasserre, 2010, Theorem 3.11)**.** *Let $\mathbf{y}_{2r} = \{y_\alpha\}_{\alpha \in \mathbb{N}^m_{2r}}$. Let $\bar{r} = \max_{1 \leq j' \leq J+1} r_{j'}$. Then the sequence $\mathbf{y}_{2r}$ has a rank$(M_r(\mathbf{y}_{2r}))$-atomic representing measure whose support is contained in $\mathbb{K}$ if and only if the following conditions hold:*

- *$M_r(\mathbf{y}_{2r})$ and $M_{r-\bar{r}}(g_j \mathbf{y}_{2r})$, $j = 1, \ldots, J+1$, are PSD,*
- *rank$(M_r(\mathbf{y}_{2r})) =$ rank$(M_{r-\bar{r}}(\mathbf{y}_{2r}))$.*

The theorem establishes the following algorithm for solving constrained polynomial optimization problems.

1. Set $r = \lfloor (d+1)/2 \rfloor$.
2. Solve (62) and compute V_r and $\mathbf{y}_{2r}$.
3. Check if $\operatorname{rank}(M_r(\mathbf{y}_{2r})) = \operatorname{rank}(M_{r-\bar{r}}(\mathbf{y}_{2r}))$.
4. If [3.] is true, then V_r is the exact minimum of the polynomial and the number of minimizers equals to $\operatorname{rank}(M_r(\mathbf{y}_{2r}))$. If [3.] is false, increase r by 1 and go to [2.].

In practice, we specify an upper bound r_0 on r and stop iteration when r reaches r_0. In that case, V_{r_0} is a lower bound for V.

C.4 Unconstrained polynomial optimization

If $\mathbb{K} = \mathbb{R}^m$, then Putinar's Positivstellensatz does not apply, but we can still use the necessary conditions to derive a SDP:

$$V_r^* = \min_{\mathbf{y}_{2r}} \sum_{|\alpha| \leq 2r} c_\alpha y_\alpha \qquad \text{subject to} \qquad \begin{aligned} & y_0 = 1, \\ & M_r(\mathbf{y}_{2r}) \text{ is PSD.} \end{aligned} \tag{63}$$

The constraint of (63) is a necessary condition for $\mathbf{y}_{2r}$ having a representing probability measure. This means that V_r^* is a lower bound for the minimum of the polynomial $f = \sum_{|\alpha| \leq 2r} c_\alpha y_\alpha$. The following theorem provides the procedure for checking whether V_r^* is actually an exact optimum.

Theorem 7 (Lasserre, 2010, Theorem 3.7)**.** *Let* $\mathbf{y}_{2r} = \{y_\alpha\}_{\alpha \in \mathbb{N}^m_{2r}}$. *Then the sequence* $\mathbf{y}_{2r}$ *has a* $\text{rank}(M_r(\mathbf{y}_{2r}))$*-atomic representing measure on* $\mathbb{R}^m$ *if and only if the following conditions hold:*

- $M_r(\mathbf{y}_{2r})$ *is PSD,*
- $M_r(\mathbf{y}_{2r})$ *admits a flat extension* $M_{r+1}(\mathbf{y}_{2r+2})$.

Then we have the following algorithm for solving unconstrained polynomial optimization problem.

1. If d is odd, then the minimum is negative infinity. If d is even, set $r = d/2$.
2. Solve (63) and compute V_r^* and $\mathbf{y}_{2r}$.
3. If $r = 2$ and $\text{rank}(M_r(\mathbf{y}_{2r})) \leq 6$ or if $r \geq 3$ and $\text{rank}(M_r(\mathbf{y}_{2r})) \leq 3r - 3$, then V_r^* is the exact minimum of f and the number of minimizers equals to $\text{rank}(M_r(\mathbf{y}_{2r}))$.
4. If $\text{rank}(M_r(\mathbf{y}_{2r})) = \text{rank}(M_{r-1}(\mathbf{y}_{2r-2}))$, then V_r^* is the exact minimum of f and the number of minimizers equals to $\text{rank}(M_r(\mathbf{y}_{2r}))$.
5. Otherwise, V_r^* is a lower bound for the minimum of f.

Step [3.] is an additional condition stated in Lasserre (2015, Theorem 2.36), which is easier to check than the flat extension condition in Step [4.].

The above algorithm solves only one SDP which do not guarantee exact solution. Nie, Demmel, and Sturmfels (2006) proposed a refinement of the algorithm which solves a sequence of the SDPs that has finite convergence to the exact solution. The idea is that the first order derivatives of a polynomial are also polynomials, and so the first order condition

$$\frac{\partial f}{v_k} = 0, \quad k = 1, \ldots, m,$$

is a polynomial constraint. In addition, the first order condition generalizes to:

$$g(v)\frac{\partial f}{v_k} = 0, \quad k = 1, \ldots, m, \tag{64}$$

for any polynomial $g(v)$, which is also a polynomial constraint. Nie, Demmel, and Sturmfels (2006) obtain a sequence of the SDPs, indexed by $r \geq \lfloor (d+1)/2 \rfloor$, by imposing (64) as additional constraints for (63). They show that the SDP sequence has finite convergence to the exact solution under suitable conditions. My R package optpoly implements this algorithm for unconstrained polynomial optimization.

C.5 Extraction of the minimizers

If the SDP satisfies certificate of optimality, we can extract minimizers of the polynomial from the optimal moment sequence $\mathbf{y}_{2r}$. Refer to Lasserre (2010, 2015) for the details of the algorithm. My R package optpoly implements the algorithm. The Gloptipoly package for Matlab also implements it.

The extraction algorithm requires that the number of minimizers is known, which equals to $\text{rank}(M_r(\mathbf{y}_{2r}))$. The rank of this matrix can be checked numerically by counting zero eigenvalues. A prior knowledge on the number of minimizers can also be useful. For example, in dynamic random coefficient models, the polynomial coefficients are drawn from a continuous distribution, in which case the polynomial has a unique minimizer with probability one.

When the polynomial has a unique minimizer, the extraction algorithm becomes simple, which I introduce here. The minimizer $v^* = (v_1^*, \ldots, v_m^*) = \text{argmin}_v f(v)$ is given by:

$$v_k^* = y_{\mathbf{e}_k}$$

where $\mathbf{e}_k$ is the vector with one at the k-th entry and zero elsewhere. The reasoning is the following. If there is unique minimizer, then the representing measure of $\mathbf{y}_{2r}$ is a point-mass distribution concentrated at the minimizer of f. This means that the moments with respect to P^* are deterministic functions of the minimizer, and in particular:

$$y_{\mathbf{e}_k} = \int v^{\mathbf{e}_k} dP^* = \int v_k dP^* = v_k^*.$$

That is, the first-order moments of P^* are coordinates of the minimizer. So the vector of the first-order moments from $\mathbf{y}_{2r}$ is the unique minimizer of f.

D Additional empirical results

D.1 Simulation

Section 5.2 mentioned that simulation results suggest confidence intervals produced by two-step procedure of Chernozhukov, Chetverikov, and Kato (2019) have coverage rate less than the nominal rate. This subsection introduces simulation results on two-step procedures.

I compute confidence intervals under the same setup as Section 7, except that the two-step procedure is used with $\beta = 0.1\alpha = 0.09$. Computational cost of two-step procedure is much higher than that of one-step because I have to rely on grid search to find confidence limits. To save computational cost, I employ another two-step approach for finding confidence limits: I first perform grid search

D.2 Application

Section 3 mentioned that Assumption 2 may be a strong assumption and its sensitivity will be checked during application by trimming observations with small eigenvalues of $\sum_{t=1}^{T} R_{it} R_{it}'$. This subsection presents results of it.