Identification and estimation of dynamic random coefficient models*

Wooyong Lee†

November 14, 2019
(click here to view the latest version)

Abstract

This paper studies dynamic panel data linear models that allow for multiplicative and additive heterogeneity in a short panel setting, by allowing both the coefficients and the intercept to be individual-specific. I show that the model is not point-identified and yet partially identified, and I characterize the sharp identified sets of the mean, variance and distribution itself of the partial effect distribution. The characterization applies to both discrete and continuous data. A computationally feasible estimation and inference procedure is proposed, which is based on a fast and exact global polynomial optimization algorithm. The method is applied to study life-cycle earnings and consumption dynamics of U.S. households in the Panel Study of Income Dynamics (PSID) dataset. The estimation results suggest that there is large heterogeneity in earnings persistence and the earnings elasticity of consumption and that there is a strong correlation between the two. Calibration of a life-cycle model suggests that heterogeneity in asset-related factors such as interest rate or discount rate is required to accurately describe real-world consumption and savings behavior.

*I am deeply indebted to Stéphane Bonhomme, Alexander Torgovitsky and Guillaume Pouliot for their invaluable 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 their helpful discussions and comments. I also thank seminar participants of Econometrics Workshop and Econometrics Working Group at University of Chicago.

†Department of Economics, University of Chicago. Email: wooyong@uchicago.edu

1 Introduction

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

In addition to the unobserved heterogeneity in levels, there is ample evidence on individuals having unobserved heterogeneity that interacts with observable characteristics. For example, firms may have different levels of efficiently in using labor and capital, households may have different levels of persistence in their earnings with respect to their past earnings, and individuals may have different levels of return to education. Such multiplicative heterogeneity is a key mechanism for heterogeneous response to exogenous shocks and policies, such as employment subsidy, income tax reform and tuition subsidy. In addition, taking account of multiplicative heterogeneity has a first order impact on more complicated models. For example, heterogeneity in the earnings persistence governs heterogeneity in the earnings risk faced by households, which is a fundamental motive for precautionary savings in the life-cycle model of consumption and savings behavior.

This paper studies the dynamic panel data linear model that allows for both multiplicative and additive unobserved heterogeneity, i.e. the dynamic random coefficient model, in a short panel setting. To be concrete, 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 to the current history of Y_{it} (up to $t-1$) but correlated to 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. In addition, the model allows the lagged outcome $Y_{i,t-1}$ to be a regressor, reflecting dynamics. Analysis of this model is challenging in short panels, since it is not possible to learn about individual values of the β's with small number of waves. This paper is the first to propose general methods for identifying and estimating moments and distributions of the β's.

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

that future values of regressors are not correlated to current outcome, which can be hard to justify. For example, a firm's labor purchase decision next year may be correlated with this year's output since the firm may learn about its own efficiency of labor from the output. Moreover, a researcher may 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 to be lasting, which reduces household's consumption smoothing ability and hence household welfare.

For random coefficient models with short panels in a dynamic setting, there are only a limited set of results. In an important paper, Chamberlain (1993) showed that the mean of β's in dynamic random coefficient models is not point-identified, which implies that the mean of β's is not consistently estimable. Arellano and Bonhomme (2012) showed that, when the regressors are binary, mean of β's for some subpopulation is identifiable and hence consistently estimable, but they did not provide a general identification result that allow consistent estimation and inference.

This paper is the first to present a general identification result for dynamic random coefficient models that allow consistent estimation and inference. Identification results for various features of β's are presented, including the mean, variance and CDF of β's. Moreover, this paper proposes a computationally feasible method for estimation and inference about the features of β's, a key step of which is to use a fast and exact algorithm for solving global polynomial optimization problems. Then the estimation method is applied to learn about heterogeneity in life-cycle earnings and consumption dynamics across U.S. households in the Panel Study of Income Dynamics (PSID) dataset. The results of this paper are presented in three steps, which are described in what follows.

First, this paper shows that dynamic random coefficient models are partially identified, which implies that there are finite bounds that can be placed on the parameters of interest. The results are general in the sense that they allow the regressors and the 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 the regressors or the coefficients are continuous. Then I use the dual representation of infinite-dimensional linear programming (Galichon and Henry, 2009; Schennach, 2014) to obtain sharp bounds for the 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 the sharp bounds obtained from the dual representation involve solving a nested optimization problem where a researcher maximizes an ob-

jective function which contains another minimization problem. An important computational issue faced here is that the inner minimization problem is a global minimization problem of a non-convex function, for which standard global optimization procedures are infeasible as 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. Then I use a fast and exact algorithm for solving global polynomial optimization problem, namely the semidefinite relaxation approach (Lasserre, 2010, 2015). Thanks to this algorithm, sharp bounds for parameters of interests can be computed in a reasonable time. Moreover, inference 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 for global polynomial optimization, I offer a general-purpose R package optpoly that implements the approach[1].

Third, I estimate a reduced-form life-cycle model of earnings and consumption dynamics and find large heterogeneity in the dynamics. The model is estimated using the Panel Study of Income Dynamics (PSID) dataset, which contains earnings, consumption and asset holdings data of U.S. households. Heterogeneity and dynamics are essential features of these data. Households may have different earnings and consumption behavior due to difference in their structural parameters such as earnings persistence or discount rate. In addition, there is dynamics in the data since past and future values of earnings, consumption and asset holdings are interrelated through the intertemporal budget constraint.

Empirical research on life-cycle earnings and consumption behavior usually assumes no heterogeneity or observable heterogeneity in the earnings persistence (Hall and Mishkin, 1982; Blundell, Pistaferri, and Preston, 2008; Blundell, Pistaferri, and Saporta-Eksten, 2016; Arellano, Blundell, and Bonhomme, 2017). This paper investigates unobserved heterogeneity in household earnings and consumption behavior as in Alan, Browning, and Ejrnæs (2018). I find that there is large heterogeneity in the earnings elasticity of consumption, i.e. household's consumption response to exogenous changes in earnings, and that the elasticity is higher when the household has higher earnings persistence. A structural model is calibrated to assess importance of unobserved heterogeneity (Kaplan and Violante, 2010; Blundell, Low, and Preston, 2013), and it is shown that heterogeneity in the earnings persistence is essential for accurately reflecting large heterogeneity in the earnings elasticity in the PSID dataset. Moreover, the calibration results suggest that heterogeneity in asset-related factors such as interest rate or discount rate is also required to accurately reflect real-world household consumption

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

behavior.

The results of this paper extend to generalized method of moments (GMM) with unobservable quantities, which can be used to answer wide range of economic questions. For example, it can be applied to the analysis of heterogeneous relationship between earnings and labor supply (Abowd and Card, 1989), or to production function estimation with firm-specific efficiency in labor and capital (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg, Caves, and Frazer, 2015).

The rest of the paper is structured as follows. From Section 2 to Section 4, a dynamic random coefficient model is formally introduced and theoretical results about the model are presented. From Section 5 to Section 6, estimation and computation method for the parameters of interest is introduced, and statistical properties of the estimator are studied. In Section 7, the method is applied to life-cycle earnings and consumption dynamics. Section 8 concludes.

2 Model and motivating examples

The dynamic random coefficient model is specified as

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

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

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

In particular, it assumes that the error term is mean independent of full history of $\{Z_{is}\}_{s=1}^T$ but of current history $\{X_{is}\}_{s=1}^t$. In other words, 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 setting, 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 same dimensions as (Z_{it}, X_{it}), and they can be freely correlated among themselves and to the observed data. The random coefficients 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.

The following simplifying 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 relation $\varepsilon_{it} = Y_{it} - Z_{it}'\gamma_i - X_{it}'\beta_i$.

This paper considers a parameter of interest 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. For theoretical results, m can be a generic Borel measurable function, but in the computation I will focus on the case where m is either a polynomial or an indicator function of V_i and propose a computationally feasible procedure for estimation and inference. Such choice of m includes many important parameters of interest. For example, θ can be an element of the mean vector $\mathbb{E}(\beta_i)$ or an element of $\mathbb{E}(\beta_i\beta_i')$. θ may 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 the indicator function $\mathbf{1}(\beta_i \leq b)$ for some b, in which case the parameter of interest 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 example 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 life-cycle earnings process, with Y_{it} being the log-earnings net of demographic variables, which is a key input for the life-cycle model of consumption and savings behavior[2]. Specification of the earnings process has a first order impact on the model outcome. Persistence of the earnings (β_i) govern earnings risk faced by households, which is a fundamental motive for precautionary savings.

The literature usually modeled 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 $\beta_i = 1$. Notable studies that allow for coefficient heterogeneity are Browning, Ejrnaes, and Alvarez (2010) and Alan, Browning, and Ejrnæs (2018).

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

Example 2 (Household consumption behavior)**.** Consider the following model of life-cycle consumption behavior:

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

where all variables are scalar, 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 this model, Y_{it} may be taken as strictly exogenous, meaning that the future earnings stream is not affected by current consumption choice. However, A_{it} must be taken as sequentially exogenous since the past and future assets and consumptions are interrelated through the intertemporal budget constraint.

(4) can be considered as an approximation to the consumption rule derived from a structural model (Blundell, Pistaferri, and Saporta-Eksten, 2016). One of the key parameters of interest in (4) is γ_{i1}, which is the elasticity of consumption to the earnings. This quantity measures household's ability to smooth consumption against exogenous changes in earnings such as exogenous earnings shocks, which is a key determinant of household's consumption smoothing ability and hence household welfare. Similar to the case of Example 1, the literature focused on models with no coefficient heterogeneity[3]

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

In the application, using data on U.S. households from Panel Study of Income Dynamics (PSID) dataset, I find that there is large heterogeneity in the elasticity of consumption to earnings (γ_{i1}) and asset holdings (β_i).

Results of this paper also extend to 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 that

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

which is a multivariate extension of (2).

[3]See e.g. Jappelli and Pistaferri (2010) for a survey.

Example 3 (Joint model of household earnings and consumption behavior)**.** One may combine (3) and (4) in Examples 1 and 2 and consider a joint life-cycle model of earnings and consumption behavior. If we combine the time t consumption equation and the time $t+1$ earnings equation, then we obtain the following 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 the following matrix form:

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

In the application, I estimate this model and study correlation between earnings persistence and the elasticities of consumption using the PSID dataset. I find that earnings persistence is strongly correlated to the earnings elasticity of consumption but not correlated to the assets elasticity of consumption. This suggests that heterogeneity in the assets elasticity of consumption stems from factors other than earnings persistence, such as heterogeneity in discount rate or interest rate.

3 Identification of means

In this and the next sections, I present theoretical results about identification of the dynamic random coefficient model defined in (1) and (2). This section focuses on identification of the means of random coefficients, and the next section presents a general identification result. Focusing on the mean allows to explain intuition of the results using simple algebra.

In this section, I consider identification of the 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 vectors that the researcher chooses and $e = (e_\gamma', e_\beta')'$. For example, one may take $e_\gamma = 0$ and $e_\beta = (1, 0, \ldots, 0)'$, in which case μ_e is the expectation of the first entry of β_i.

Identification results about μ_e are presented in three subsections. In the first subsection, μ_e is shown to be not point-identified in general. Then the following subsection shows that μ_e is

partially identified. The third subsection shows that conditioning on (γ_i, β_i) in (2) is essential for partial identification.

3.1 Failure of point-identification

This subsection shows that μ_e is not point-identified in general. This is done 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 shows that $\mathbb{E}(\beta_i)$ is not point-identified in this model.

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 they 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 sufficiently large relative to that of (Y_{i0}, Y_{i1}, Y_{i2}). The proof suggests that the result holds for any finite $T \geq 2$. Note that the failure of point-identification in Proposition 1 implies that there is no consistent estimator for $\mathbb{E}(\beta_i)$.

The proof for Proposition 1 consists of two steps. First, I show that $\mathbb{E}(\beta_i)$ is point-identified if and only if there exists an estimator $S(Y_{i0}, Y_{i1}, Y_{i2})$ from the individual time series that is unbiased for β_i[4]. Then it is shown that there is no unbiased estimator for β_i. The intuition for the first step is that, since the distribution of β_i is unrestricted and hence β_i can take any value within $\mathcal{C}$, information on β_i for a given i can only be obtained from individual i's data (Y_{i0}, Y_{i1}, Y_{i2}). If the individual data provides an "exact" information about β_i in the sense of unbiasedness, then we obtain point-identification of $\mathbb{E}(\beta_i)$. However, if there is no unbiased information, we do not achieve point-identification.

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} only. Proposition 1 generalizes his result and

[4]This result also holds for a general case considered in the next section.

shows that point-identification is also not possible with stronger assumption and continuous data. 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.

Then a natural question following Proposition 1 is whether the data contain any information about $\mathbb{E}(\beta_i)$ or there is no information at all. The next subsection shows that the data are actually informative about $\mathbb{E}(\beta_i)$. More precisely, it is shown that μ_e is partially identified for any fixed e in (1).

3.2 Partial identification

In what follows, I show that μ_e is partially identified under suitable conditions. In particular, I show that there are finite bounds L and U such that

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

where L and U are estimable with data. This implies that there exist consistent estimators for lower and upper bounds of μ_e. The quantities L and U depend on e, but the dependence is suppressed in the notation.

Using the notation defined in Section 2 and letting $R_{it} = (Z_{it}', X_{it}')'$ be the vector of regressors at time t, let's 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) and characterize the sharp identified set under those restrictions. 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). In this paper, I choose the class of g to be the set of polynomials in its arguments and use a finite subset of them for estimation and inference. The finite subset of unconditional moment restrictions contains less information than (7), but it yields a computationally feasible estimation and inference procedure for the parameters of interest. Partial identification results based on (7) can be derived using Theorem 3 in Appendix B.

In this section, I consider the following assumptions.

Assumption 1. The 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. The 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, which states that there is variation in the regressors over time for every individual. Note that the dynamic fixed effect model only requires that the expectation $\mathbb{E}(\sum_{t=1}^T R_{it} R_{it}')$ is positive definite in order for the data to be informative about the model. Here I require the assumption for each individual since V_i is individual-specific and I require information for all i. Without Assumption 2, there may be individuals with no information about V_i in the data, and their V_i values may be arbitrarily large or small so that we cannot learn about $\mathbb{E}(e' V_i)$. Although it is necessary, it can be a strong assumption in empirical settings, and in the application I trim observations with small eigenvalues of $\sum_{t=1}^T R_{it} R_{it}'$ and check sensitivity of my results against Assumption 2.

Equations in Assumption 3 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 full history of Z_{it} and 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 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*

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

and

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

where

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

Moreover, these are the sharp bounds of μ_e under Assumptions 1 and 3.

Proof. See Appendix A.2. □

Note that, 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). However, as mentioned earlier, L and U in Theorem 1 imply computationally feasible estimators, which I exactly compute in the application.

Theorem 1 characterizes the lower and upper bounds of μ_e as solutions to optimization problems over the Euclidean space. Dimension of this space can be potentially large. However, it will be shown that the optimization can be performed fast and reliably.

L and U have closed-form expressions, but I do not display them here since (i) they are very complicated and (ii) they can be computationally more demanding than solving the optimization problems as they 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 that 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. In addition, $[\tilde{L}, \tilde{U}]$ are sharp bounds of μ_e under Assumptions 1 and 2 and the following implication of Assumption 3:

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

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 the bounds are finite as long as the moments about $\mathcal{R}_i$ and $\mathcal{Y}_i$ are finite, namely $\mathbb{E}(\mathcal{R}_i)$, $\mathbb{E}(\mathcal{Y}_i)$, $\mathbb{E}(\mathcal{R}_i^{-1}\mathcal{Y}_i)$ and $\mathbb{E}(\mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i)$. Note that $\mathcal{R}_i$ 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.

Now I explain the intution behind Theorem 1. Let's focus on the upper bound U. For any given (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. One may interpret Q as "Lagrangian": it is a linear combination of $e'V_i$ and the moment functions with Lagrange multipliers $\{\lambda_t\}$ and $\{\mu_t\}$. Note that $\mathbb{E}(Q) = \mathbb{E}(e'V_i)$ since the second and the third terms have zero expectation by Assumption 3.

If we substitute $\varepsilon_{it} = Y_{it} - R_{it}V_i$ into Q, we obtain the following 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 has 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". Then

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

Taking 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 $\lambda > 0$. Moreover, since the above equation holds for any (λ, μ) such that $\lambda > 0$, it follows that

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

which is shown to be the sharp upper bound in the proof of Theorem 1. The sharp lower bound can be obtained by repeating the same argument with $\lambda < 0$.

3.3 Identifying power of the standard assumption in fixed effect models

In dynamic fixed effect models, a frequently used condition is that the error term is mean independent of (Z_i, X_i^t) but not necessarily V_i. In other words, the following assumption is often made:

Assumption 4. The random variables $(W_i, V_i)_{t=1}^T$ and $(\varepsilon_{it})_{t=1}^T$ satisfy

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

Note that V_i is not included as a conditioning variable in Assumption 4. In standard fixed effects models with no coefficient heterogeneity, Assumption 4 is sufficient for identifying and estimating the coefficients.

It can be shown that Assumption 4 alone provides no information at all about the coefficients, which is stated as the following proposition.

Proposition 3. *Suppose Assumptions 1 and 4 hold. Suppose that $e \neq 0$. Consider a sequence of distributions P_M, indexed by $M \in \mathbb{N}$, such that the support of (W_i, V_i) is $[-M, M]^{(1+q+p)T+(q+p)}$.*

Let $[L_M, U_M]$ be the sharp bound of μ_e. Then, under regularity conditions on $\{P_M\}$,

$$\lim_{M\to\infty} L_M = -\infty \quad \textit{and} \quad \lim_{M\to\infty} U_M = \infty.$$

That is, as $M \to \infty$, the identified set tends to a trivial set.

Proof. See Appendix A.4. □

Note that Chamberlain (1993) showed failure of point-identification under Assumptions 1 and 4. Proposition 3 shows the size of the identified set under the assumption.

This result is related to the result of Ahn and Schmidt (1995) where they point out that additional moment conditions other than Assumption 4 can provide additional information about the model. Proposition 3 is an extreme case where Assumption 4 provides no information at all without additional moment conditions.

The intuition for Proposition 3 is similar to that of Theorem 1. For any given function $g_t : \mathbb{R}^{qT+pt} \mapsto \mathbb{R}$, consider the quantity

$$\tilde{Q}(g_1, \ldots, g_T, W_i, V_i) = e'V_i + \sum_{t=1}^{T} g_t(Z_i, X_i^t)\varepsilon_{it}.$$

Note that $\mathbb{E}(\tilde{Q}) = \mathbb{E}(e'V_i)$ since the second term has zero expectation by Assumption 4.

If we substitute $\varepsilon_{it} = Y_{it} - R_{it}V_i$ into $\tilde{Q}$, we can see that $\tilde{Q}$ is a linear function of V_i:

$$\tilde{Q}(g_1, \ldots, g_T, W_i, V_i) = \sum_{t=1}^{T} g_t(Z_i, X_i^t)Y_{it} + \left[e - \sum_{t=1}^{T} g_t(Z_i, X_i^t)R_{it}\right]' V_i.$$

Since a linear function is unbounded and $e \neq 0$, the maximum of $\tilde{Q}$ with respect to V_i is infinite for any g_t, except for a probability-zero set of W_i. Then it follows that the upper bound of μ_e, given by the expectation of the maximum of $\tilde{Q}$ with respect to V_i, is infinite. The proof of Proposition 3 establishes a formal argument of this intuition.

4 Identification of higher order moments and the CDFs

This section formalizes the intution of Section 3 and presents a general partial identification result for dynamic random coefficient models. 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 the following unconditional moment restrictions.

Assumption 5. The random vectors (W_i, V_i) satisfy

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

where the ϕ_k's are real-valued moment functions and K is the number of moment restrictions.

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 where the ϕ_k's for $k = 1, \ldots, T$ are

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

and the ϕ_k's 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.

Note that the error term ε_{it} does not appear in Assumption 5 since ε_{it} is understood as a deterministic function of (W_i, V_i) by the relation $\varepsilon_{it} = Y_{it} - R_{it}'V_i$.

Now I characterize the identified set of θ under Assumptions 1 and 5. The approach is to directly write down the definition of the identified set and then characterize it. 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. Also, 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

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

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

In words, I is the collection of all $\int m(W_i, V_i)dP$ values that are 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 to the observed distribution P_W.

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

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

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

Equation (9) is a linear program (LP) in P, with a 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 is computationally infeasible due to the curse of dimensionality. 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 since the dimension of (W_i, V_i) is often large. Note that W_i contains full history of all observables (regressors and response variables) and that 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.

The approach employed here is to use the dual representation of (9). The standard duality theorem for finite-dimensional LP extends to the infinite-dimensional case, and we can use the dual representation of (9) to obtain a tractable characterization of I.

The following theorem characterizes I using the dual representation of (9).

Theorem 2. *Suppose Assumption 5 hold. 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.5. □

Note that Assumption 1 is not included in Theorem 2. The result of Theorem 2 applies more generally to models of generalized method of moments (GMM) where the moment functions contain both observables and unobservables (Galichon and Henry, 2009; Schennach, 2014).

Although Theorem 2 applies to models of GMM, the estimators that Theorem 2 imply are not obvious to compute for general models of GMM. In the next section, I show that, for dynamic random coefficient models, the linear structure of the model can be exploited to obtain a computationally tractable estimation procedure.

5 Computation

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 the expectations in (10) and (11) with the sample means. Define $\hat{L}$ as an estimator for L where

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

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

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

Statistical properties of (12) and (13) are studied in the next section. This section discusses computational issues in the estimators that are not obvious to deal with and how to resolve those issues. The discussion focuses on (12), and the same discussion applies to (13) as well.

Note first that computation of (12) requires solving two types of optimization problems, namely the inner minimization problem (for each i) and the outer maximization problem. Each problem has its own difficulties:

- the inner minimization problem in (12) must be solved globally but its objective function is not necessarily convex. Moreover, it must 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 in (12) must also be solved globally, and it may be an optimization over a large dimensional space.

In particular, the inner problem is computationally hard to deal with. 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*. For this reason, general-purpose minimization methods for the inner problem are not computationally feasible except for low-dimensional cases such as $\mathcal{V}$ is a discrete set or $\mathcal{V}$ is a convex subset of $\mathbb{R}$ or $\mathbb{R}^2$.

The results about computation are presented in two subsections. The first subsection shows that, for random coefficient models, the inner problem can be solved fast and exactly by using a fast and exact algorithm for global polynomial optimization. Then the second subsection 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.

5.1 The inner problem

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

$$G(\lambda_1, \ldots, \lambda_K, w) = \min_{v \in \mathcal{V}} \left\{ m(w, v) + \sum_{k=1}^{K} \lambda_k \phi_k(w, v) \right\} \tag{14}$$

for each fixed $w = W_i$, where $i = 1, \ldots, N$, given the value of $(\lambda_1, \ldots, \lambda_K)$.

A difficulty in evaluating G is that the minimization problem must be solved globally. In the simple case that $\mathcal{V}$ is discrete or is a low-dimensional space 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, the assumption that the random coefficients are discrete is often hard to justify, and the dimension of $\mathcal{V}$ is often larger than two. Note that the dimension of $\mathcal{V}$ equals to the number of regressors in dynamic random coefficient models, including a constant.

This subsection shows that G can be computed fast and exactly when m and ϕ_k's are polynomials in v. This is because, when m and the ϕ_k's are polynomials, evaluation of G is equivalent to solving global minimization problem of a polynomial, for which a fast and exact algorithm exists. Note that we can choose ϕ_k's to be polynomials since ε_{it} is a linear function of V_i and we can choose g to be polynomials in V_i in (8).

The polynomial case is useful for 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. Moreover, the ϕ_k's under Assumption 3 consist of the functions

$$(R_{it}'V_i)(Y_{it} - R_{it}'V_i), \quad t = 1, \ldots, T, \tag{15}$$

and the entries of the vectors

$$(Z_i', X_i^{t\prime})'(Y_i - R_{it}'V_i), \quad t = 1, \ldots, T, \tag{16}$$

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

Example 6. Suppose a researcher is interested in identifying an element of $\mathbb{E}(V_iV_i')$. In this case, the researcher sets m to be 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_k's 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 may also assume that Assumption 3 holds, in which case one sets the additional ϕ_k's to be (15) and (16). These moment restrictions are exactly what I use in the application for 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 for ensuring finite bounds, a researcher can choose ϕ_k's so that the inner objective function has even order that is strictly larger than the order of the parameter of interest. In Examples 5 and 6, I chose ϕ_k's to be the second order for $\mathbb{E}(V_i)$ and to be the fourth order for $\mathbb{E}(V_iV_i')$. Then the inner objective function 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_k's to be indicator functions of V_i. The idea is that an indicator function partitions $\mathcal{V}$ into two exclusive sets and that the

indicator function is constant within each set. Then the researcher can compute the global optimum within each partition and then compute the optimum of the two.

This extension is useful for computing bounds for CDFs of random coefficients, which is described as 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. Then 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. Let's assume that Assumption 3 holds, in which case the ϕ_k's are at most second-order polynomials in V_i stated in (16) and (15). Then 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 $\mathcal{V}_2 = \{(v_1, \ldots, v_{q+p}) \mid v_1 > v\}$. In addition, $m = 1$ on $\mathcal{V}_1$ and $m = 0$ on $\mathcal{V}_2$. Then the objective function in (14) is a second-order polynomial within each of $\mathcal{V}_1$ and $\mathcal{V}_2$ for which one can compute the minimum. Then the researcher can evaluate G by taking the smaller optimum between those in $\mathcal{V}_1$ and $\mathcal{V}_2$.

The following two subsections discuss a fast and exact computation method for global optimization of polynomials. The first subsection considers a simple case of quadratic polynomials for which the global solution is obtained in a closed-form. Then the second subsection considers generic polynomials for which the global optimization problem is solved numerically.

5.1.1 Global optimization of quadratic polynomials

We first consider a simple case of global optimization of quadratic polynomials. We can express a quadratic polynomial in the following 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 $c \in \mathbb{R}$. Note that, if (14) is expressed in this standard form, (A, b, c) are functions of w.

The first and the second derivatives of $Q(v)$ are

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

The global optimum of Q can be computed using simple algebra. First, if A is positive definite,

Q is globally convex and it 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 given by

$$\min_{v \in \mathcal{V}} Q(v) = c - \frac{1}{4} b' A^{-1} b. \tag{17}$$

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

$$2Av + b = 0$$

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

In practice, when we solve (14) for each $w = W_i$ and if W_i follows a continuous distribution, A possesses a zero eigenvalue with probability zero. Therefore, we may simply use (17) to express (14) in a closed-form if and only if A is positive definite; otherwise we assign negative infinity.

5.1.2 Global optimization of generic polynomials

When m and ϕ_k's are polynomials of generic order, a closed-form solution is not avilable, 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 moreover it computes an *exact* solution. This subsection discusses main idea of the algorithm. A formal discussion can be found in 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 $p_j(v)$. We can express a fourth-order polynomial

in the following standard form:

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

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

$$\min_{P_V \in \mathcal{M}_V,\ \int dP_V = 1} \int \pi(v) dP_V \tag{18}$$

where P_V is a probability distribution on $\mathcal{V}$. (18) is minimized at the point-mass distribution concentrated at the minimizer of $\pi(v)$.

Then, since $\pi(v)$ is a linear combination of $p_j(v)$, we can rewrite (18) 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

$$\min_{M_1,\ldots,M_J \in \mathbb{R},\ M_1 = 1} \sum_{j=1}^{J} a_j M_j \qquad \text{subject to} \qquad M_j = \int p_j(v) dP_V \text{ for some } P_V \in \mathcal{M}_V. \tag{19}$$

Except for the fact that the constraint is complicated, (19) is a minimization over $\mathbb{R}^J$ and the objective is linear (and hence convex) in the choice variables.

Now the idea is to replace the constraint in (19) with a convex constraint that only involves $(M_1, \ldots, M_J)$. The constraint in (19) tells that $(M_1, \ldots, M_J)$ must be moments of some underlying distribution. Checking this constraint is related to the following 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 relations between them in order for them to be justified as moments. For example, for a generic real random variable X, it must be that $\mathrm{Var}(X)$ is positive, i.e.

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

This is equivalent to the following 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 a linear operator $\mathcal{L}$ that maps a polynomial to $\mathbb{R}$ by the relation

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

It can be shown that, if $(M_1, \ldots, M_J)$ are moments, then

$$\mathcal{L}(u(v)u(v)') \text{ is positive semidefinite} \tag{20}$$

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

We can show that (20) is a convex constraint. It is based on the fact that the set of positive semidefinite matrices is a convex set in the space of vectorized matrix entries. Therefore, if we replace the constraint in (19) with (20), we obtain the following convex optimization problem:

$$\min_{M_1,\ldots,M_J \in \mathbb{R}} \sum_{j=1}^{J} a_j M_j \qquad \text{subject to} \qquad \mathcal{L}(u(v)u(v)') \text{ is positive semidefinite.} \tag{21}$$

Moreover, the constraint can be handled more efficiently than a generic convex constraint so that the optimization problem has its own name: semidefinite program (SDP). It is an optimization problem where a matrix involving the choice variables is constrained to be positive semidefinite.

The SDP approach of polynomial optimization solves (21), the *semidefinite relaxation*, which can be solved fast and reliably using SDP solvers available in the industry. The algorithm offers *certificate* of optimality, a sufficient condition in terms of the optimal value of $(M_1, \ldots, M_J)$ which ensures that the solution to (21) equals to the exact global optimum. For researchers interested in using the semidefinite relaxation approach for global polynomial optimization, I offer a general-purpose R package `optpoly` which implements the approach[5]. Alternatively, another 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 (21), i.e. the SDP solution, is less than or equal to the solution to (19). The semidefinite relaxation approach solves a hierarchy of the SDP programs, or a *sequence* of the SDP programs, until one obtains the certificate of optimality which is known to be obtained in finite number of steps under suitable conditions. Moreover, even if one does not solve the hierarchy of the

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

SDPs, one can take an SDP solution as a lower bound for (19) and the resulting value of (12) is a conservative and yet a valid lower bound for θ.

5.2 The outer problem

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

$$\max_{\lambda_1,\ldots,\lambda_K} \frac{1}{N}\sum_{i=1}^{N} G(\lambda_1,\ldots,\lambda_K, W_i).$$

Assume that the researcher can evaluate G exactly using the algorithm in the previous subsection. Then the remaining difficulty 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.

Proposition 4. *Let $\lambda = (\lambda_1,\ldots,\lambda_K)$ and define*

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

where G is defined in (14). Then $\hat{L}(\lambda)$ is globally concave in λ.

Proof. See Appendix A.6. □

Proposition 4 tells that there is only one local maximum of $\hat{L}(\lambda)$, which is also the global maximum. This suggests that the researcher can maximize $\hat{L}(\lambda)$ using fast convex optimization algorithms such as gradient descent methods. Milgrom and Segal (2002, Theorem 3) provides conditions under which G is differentiable when $K = 1$, which can be used to provide conditions under which G is directionally differentiable. In practice, if one is concerned with differentiability, one may apply gradient descent methods based on finite differences.

Note that Proposition 4 comes from the concavity of G and that solving the inner problem exactly by the polynomial optimization algorithm is crucial for computational tractability of the outer problem when K is large. This is a key distinction from the general-purpose approach of Schennach (2014) where I specifically focus on random coefficient models and exploit the structure of the model to gain 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, in the application in Section 7, K ranges from 55 to 166 depending on the model.

6 Estimation and inference

This section studies statistical properties of $\hat{L}$ defined in (12). It consists of three subsections. The first subsection shows consistency of $\hat{L}$ to L. The second subsection notes that $\hat{L}$ may not be well-defined depending on the sample, and it discusses how to modify $\hat{L}$ to be well-defined in that case. The third subsection discusses inference about L using the results in the second subsection as an input.

6.1 Consistency

Using the notation defined in (14) and Proposition 4, one can concisely write the lower bound estimator $\hat{L}$ defined in (12) as

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

Note that the inner solution function

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

does not involve any statistical object: given the model, i.e. given m and the ϕ_k's, the function G is a deterministic object. Therefore, what is studied here is the property of the statistical object

$$\hat{L} = \max_{\lambda} \hat{L}(\lambda) = \max_{\lambda} \frac{1}{N} \sum_{i=1}^{N} G(\lambda, W_i) \tag{22}$$

as an estimator for

$$L = \max_{\lambda} L(\lambda) = \max_{\lambda} \mathbb{E}\left(G(\lambda, W_i)\right). \tag{23}$$

Note that $\hat{L}(\lambda)$ is the objective function of an M-estimation problem where $L(\lambda)$ is the population objective and λ is the parameter that is M-estimated. Then the consistency follows by replicating the analysis of M-estimation. Most of the regularity conditions in M-estimation are satisfied by the fact that G is concave in λ.

Proposition 5. *Suppose that L exists and is finite and that* $\operatorname{argmax}_\lambda L(\lambda)$ *is in the interior of* $\mathbb{R}^K$. *Then $\hat{L}$ converges to L in probability.*

Proof. See Appendix A.7. □

6.2 Relaxation of moment conditions

In contrast to standard M-estimation, $\hat{L}$ may not always be well-defined. The intuition for why can be gained by comparing estimation of the random coefficient model, which is defined by moments with unobservables, to a standard GMM estimation. In the standard GMM estimation, the minimum GMM objective may be strictly positive in the sample because the moment conditions based on the empirical distribution may not be exactly satisfied due to sample variation. This also happens to random coefficient models. That is, there may be no distribution of the random coefficients that satisfy all the moment conditions given the empirical distribution of data. In this case, a researcher obtains an empty identified set, and the maximization problem of $\hat{L}$ diverges to $+\infty$ and the corresponding problem for the upper bound diverges to $-\infty$.

In standard GMM estimation, when the moment conditions are not exactly satisfied, a researcher minimizes the GMM criterion and choose the parameter value that minimizes the criterion. A similar approach can be taken here, which can be implemented in two steps[6]. In the first step, one finds the smallest $\delta \geq 0$ that satisfies the following criterion:

$$|\mathbb{E}(\phi_k(W_i, V_i))| \leq \delta, \quad k = 1, \ldots, K. \tag{24}$$

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

Proposition 6. *Given the sample* $(W_1, \ldots, W_N)$, *consider the linear programming problem*

$$\begin{aligned} \min_{P \in \mathcal{M}_{W \times V},\, P \geq 0,\, \delta \geq 0} \delta \qquad \textit{subject to} \qquad & \left| \int \phi_k(W_i, V_i) dP \right| \leq \delta, \quad k = 1, \ldots, K, \\ & \int P(w, dV_i) = \hat{P}_W(w) \ \textit{ for all } w \in \mathcal{W}, \end{aligned} \tag{25}$$

where $\hat{P}_W$ is the empirical distribution of W_i constructed from $(W_1, \ldots, W_N)$. *Then its solution equals*

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

to the solution of the following optimization problem:

$$\max_{\lambda_1,\ldots,\lambda_K} \frac{1}{N}\sum_{i=1}^{N} \min_{v\in\mathcal{V}} \left\{ \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} \qquad \textit{subject to} \qquad \sum_{k=1}^{K} |\lambda_k| \leq 1. \tag{26}$$

Proof. See Appendix A.8. □

Proposition 6 tells that a researcher can find the minimum δ by solving the problem that is similar to (12). In particular, one can use the same computation methods described in the previous section to solve the inner and the outer optimization problems in (26). A difference is that it is a constrained optimization problem, but the constraint has a very simple structure and its Jacobian can be derived in closed-form as well.

Let δ^* be the solution to (26). Then the second step computes the bounds for the parameter of interest. One computes the lower bound with the L^1 penalty on the λ's with δ^* being the penalty multiplier:

$$\hat{L}_{pen} = \max_{\lambda_1,\ldots,\lambda_K} \left[\frac{1}{N}\sum_{i=1}^{N} \min_{v\in\mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} - \delta^* \sum_{k=1}^{K} |\lambda_k| \right]. \tag{27}$$

The following proposition justifies the use of the L^1 penalty.

Proposition 7. *Given the sample $(W_1, \ldots, W_N)$ and given $\delta^* \in \mathbb{R}$, consider the linear programming problem*

$$\min_{P\in\mathcal{M}_{W\times V},\ P\geq 0} \int m(W_i, V_i) dP \qquad \textit{subject to} \qquad \begin{array}{l} \left| \int \phi_k(W_i, V_i) dP \right| \leq \delta^*, \quad k = 1, \ldots, K, \\ \int P(w, dv) = \hat{P}_W(w) \ \textit{ for all } w \in \mathcal{W}. \end{array} \tag{28}$$

where $\hat{P}_W$ is the empirical distribution of W_i constructed from $(W_1, \ldots, W_N)$. Then its solution equals to $\hat{L}_{pen}$ defined in (27).

Proof. See Appendix A.9. □

Proposition 7 tells that (27) equals to the smallest value of θ for the distributions that minimize the absolute-value GMM criterion defined in (24). In principle, such distribution is not necessarily unique. If it is unique, then the resulting estimate of the identified set from the two-step procedure becomes a point.

In practice, because of either machine precision or stopping criterion of numerical optimization methods, the numerical solution to (26) may be strictly smaller than the analytical

solution δ^*. In that case, (27) will diverge to infinity since the penalty multiplier is not large enough. To resolve this problem, one may inflate the value of numerical solution a little bit, in which case (27) 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 inflated numerical solution will be a very small interval instead of a point.

Although (27) allows the lower bound estimate to be always well-defined, there are two problems with it. First, it is an ad-hoc approach and there is no formal justification for why relaxation of moment conditions is a good idea. Second, picking up the distributions that minimize the criterion may produce a point even if the model is partially identified. The second problem can be resolved by supplying a value that is larger than δ^* in (27), but how much δ^* should be increased remains a question. The next subsection discusses a more principled approach of estimating the bounds, which is directly computing a confidence interval for the identified set.

6.3 Inference

This subsection discusses construction of a confidence interval for the identified set $[L, U]$. The objective is to compute L_α and U_α given one-sided significance level α such that

$$\liminf_{N\to\infty} \inf_{P} P([L, U] \subseteq [L_\alpha, U_\alpha]) \geq 1 - 2\alpha.$$

To compute a confidence interval, a researcher can leverage results from the literature on moment inequalities. Recall the notation in (22) and (23). The idea is to consider the lower bound L as a parameter of interest in the moment inequalities model. Define L to be the smallest number such that

$$L(\lambda) - L = \mathbb{E}\left(G(\lambda, W_i) - L\right) \leq 0 \quad \text{for all } \lambda \in \mathbb{R}^K. \tag{29}$$

This is a standard moment inequalities model albeit the number of inequalities is infinite (which is indexed by $\lambda \in \mathbb{R}^K$). To make it computationally tractable, let Λ_F be a finite subset of $\mathbb{R}^K$ and consider the following moment inequalities model with parameter L_F:

$$\mathbb{E}\left(G(\lambda, W_i) - L_F\right) \leq 0 \quad \text{for all } \lambda \in \Lambda_F. \tag{30}$$

Now the argument is that, since (30) uses smaller number of moment inequalities than (29),

one can use (30) to make a conservative inference about L.

How much conservative it is depends on how much information is contained in (30) relative to (29). Analysis on comparison between (29) and (30) is a topic of future research[7]. However, there two conjectures that can be made. First, since G is concave in λ and hence continuous in λ, one may conjecture that setting Λ_F to be a grid of $\mathbb{R}^K$ yields an arbitrary good approximation to (29). Second, concavity of G implies that there is only one moment that is binding in (29), which is the moment with the index $\lambda^* = \text{argmax}_\lambda \mathbb{E}(G(\lambda, W_i))$. The sample counterpart of λ^* can be obtained by using the estimation methods described in the previous subsections, and one can conjecture that it is sufficient to consider a grid around λ^* and there is not much information outside of its neighborhood[8].

The rest of this subsection discusses a practical approach resulting from this idea. The approach relies on the hypothesis testing method proposed by Romano, Shaikh, and Wolf (2014), which is shown to have a good performance when the size of Λ_F is large in Bai, Santos, and Shaikh (2019).

Let α be one-sided significance level and ξ be an additional tuning parameter such that $0 < \xi < \alpha$. Romano, Shaikh, and Wolf (2014) find that $\xi = 0.1\alpha$ shows a good performance in their simulation. Consider a finite grid Λ_F around λ_N^*, where λ_N^* is obtained from either the lower bound estimate in (12) or its moment-relaxation version in (27). The size of the grid can be taken to be large. For each $\lambda \in \Lambda_F$, compute

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

and

$$\hat{S}(\lambda) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(G(\lambda, W_i) - \hat{L}(\lambda)\right)^2}.$$

Note that these quantities do not involve L_F. Next, compute a critical value $c^*(\alpha, \xi)$ which also do not depend on L_F, where the details about the computation procedure is described in the appendix. Then the confidence region for L_F proposed in Romano, Shaikh, and Wolf

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

[8]This is related to a step in the inference procedure of Chernozhukov, Lee, and Rosen (2013) where they compute a set of moment restrictions that is likely to bind.

(2014) is the set of l's such that

$$\max_{\lambda \in \Lambda_F} \frac{\sqrt{N}(\hat{L}(\lambda) - l)}{\hat{S}(\lambda)} \leq c^*(\alpha, \xi).$$

This equation can be solved for l. In particular, the confidence interval for L_F is $[L_\alpha, \infty)$ where

$$L_\alpha = \max_{\lambda \in \Lambda_F} \left[\hat{L}(\lambda) - c_{\alpha,\beta} \times \frac{\hat{S}(\lambda)}{\sqrt{N}} \right].$$

Romano, Shaikh, and Wolf (2014) showed that, under suitable conditions, the confidence interval $[L_\alpha, \infty)$ is uniformly consistent in level pointwisely in the identified set:

$$\liminf_{N \to \infty} \inf_{P \in \mathcal{P}} P(L_\alpha \leq L_F) \geq 1 - \alpha,$$

where $\mathcal{P}$ is the set of distributions such that $\max_{\lambda \in \Lambda_F} \mathbb{E}_P(G(\lambda, W_i)) = L_F$. Using this result, one can show that, using $[L, U] \subseteq [L_F, U_F]$,

$$\liminf_{N \to \infty} \inf_{P \in \mathcal{P}} P([L, U] \subseteq [L_\alpha, U_\alpha]) \geq 1 - 2\alpha,$$

where U_α is the upper bound of a confidence interval for U obtained by the symmetric procedure. Note that the inequality may be strictly positive, meaning that the inference may be conservative.

7 Application to life-cycle consumption dynamics

This section applies the methods discussed so far to the panel data of life-cycle earnings and consumption dynamics. I use the data on U.S. households from Panel Study of Income Dynamics (PSID) dataset, which is sampled every two years in 1999-2015. Details about construction of the dataset can be found in the Appendix. The dataset construction procedure gives $N = 684$ individuals and $T = 8$ waves.

Application is divided into two subsections. The first subsection estimates a reduced-form model of life-cycle earnings and consumption dynamics. The second subsection calibrates a structural life-cycle model to make sense of the reduced-form estimates.

7.1 Reduced-form estimation

Consider the reduced-form model of life-cycle earnings and consumption dynamics

$$\begin{aligned} C_{it} &= \gamma_{i0} + \gamma_{iY} Y_{it} + \gamma_{iA} A_{it} + \nu_{it}, \\ Y_{i,t+1} &= \beta_{i0} + \beta_{iY} Y_{it} + \varepsilon_{it}, \end{aligned} \tag{31}$$

where C_{it} is non-durable consumption, Y_{it} is earnings and A_{it} is asset holdings, which are all measured in logs and net of demographic variables such as education level, year-of-birth, race, etc. The full set of demographic variables are listed in the Appendix.

Let $\gamma_i = (\gamma_{i0}, \gamma_{iY}, \gamma_{iA})$ and $\beta_i = (\beta_{i0}, \beta_{iY})$, and let $Y_i = (Y_{i0}, \ldots, Y_{iT})$ be the full history of Y_{it} and $Y_i^t = (Y_{i0}, \ldots, Y_{it})$ be the current history of Y_{it} up to t. Define similar notation for C_{it}, A_{it} and Y_{it}. The error terms ν_{it} and ε_{it} in (31) satisfy mean independence assumptions

$$\begin{aligned} \mathbb{E}(\nu_{it} | \gamma_i, \beta_i, Y_i, A_i^t) &= 0, \\ \mathbb{E}(\varepsilon_{it} | \gamma_i, \beta_i, {Y_i}^t) &= 0. \end{aligned} \tag{32}$$

In particular, Y_{it} is strictly exogenous and A_{it} is sequentially exogenous in the consumption equation, and Y_{it} is sequentially exogenous in the earnings equation.

For the earnings process, it is standard in the literature to allow for two types of earnings shocks, namely permanent and transitory shocks. In (31), these earnings shocks are summarized into β_{iY}. If β_{iY} is close to 1, then the earnings process is very persistent, meaning that permanent shock dominates transitory shock. If β_{iY} is close to 0, then the earnings process is not persistent, meaning that transitory shock is a major source of earnings shock.

The consumption process in (31) is specified to be a linear function of earnings and asset holdings. While the model is a reduced-form, it can be considered as an approximation to the structural decision rule. For example, Blundell, Pistaferri, and Preston (2008) log-linearized the Euler equation of a dynamic life-cycle model and obtained a linear model of consumption decision.

γ_{iY} is a key parameter of interest in the consumption process. It measures consumption response to exogenous change in earnings, namely the earnings elasticity of consumption. It can be determined by various factors such as earnings persistence, current level of earnings, or current level of asset holdings. (31) allows to estimate the correlation between the earnings elasticity and these factors. With (31) being a random coefficient model, I impose no assumption on the correlation structure of the earnings elasticity.

γ_{iA} is another key parameter of interest, which is the asset holdings elasticity of consump-

tion. It is the partial effect of asset holdings on consumption keeping earnings *fixed*, which measures consumption response to exogenous change in asset holdings. γ_{iA} speaks about household's consumption response to fiscal policies such as fiscal stimulus payments. (31) allows to estimate the correlation between γ_{iA} and the factors such as earnings persistence, earnings elasticity of consumption, current level of earnings or current level of assets. An attactive feature of (31) is that it does not make any assumption about the evolution of the assets, which means that it allows the law of motion for assets to be nonparametric and stochastic.

In what follows, I estimate the first and the second order moments of the random coefficients. The first and the second order moments are estimated using the moment conditions described in Example 5 and Example 6, respectively. The second moments between the earnings and the consumption equations are estimated using a multivariate version of Example 6.

Then these moment estimates are used to compute the variances and the correlations. Note that the bounds on the moments are marginal bounds, which implies that the bounds on variances and correlations are non-sharp bounds. In all estimation procedures, the moment relaxation approach discussed in Section 6 is used to compute the bounds for the distributions that near-minimize the GMM criterion.

The moment estimates are presented in Table 1. For comparison, the fixed effect estimates and its bias-reduced version (Dhaene and Jochmans, 2015) is also provided. Sensitivity and robustness checks about the estimates can be found in the Appendix.

Parameter	LB	UB
$\mathbb{E}(\gamma_{iY})$	0.191	0.322
$\mathbb{E}(\gamma_{iA})$	0.153	0.212
$\mathbb{E}(\beta_{iY})$	0.287	0.425
$\mathrm{Var}(\gamma_{iY})$	0.790	1.572
$\mathrm{Var}(\gamma_{iA})$	0.155	0.308
$\mathrm{Var}(\beta_{iY})$	0.349	0.838
$\mathrm{Corr}(\gamma_{iY}, \gamma_{iA})$	-0.267	-0.008
$\mathrm{Corr}(\gamma_{iY}, \beta_{iY})$	0.042	0.846
$\mathrm{Corr}(\gamma_{iA}, \beta_{iY})$	-0.233	0.080

Table 1: Estimates of the parameters of interest. The lower bound ("LB") and the upper bound ("UB") are obtained by solving (27) and the corresponding upper bound problem with inflated δ^*.

The reduced-form estimates suggest that there is large heterogeneity in the earnings elasticity of consumption (γ_{iY}) and that its correlation to earnings persistence (β_{iY}) can be as large as 0.519. This is not captured in the fixed effects model whose correlation estimate is close to zero. In addition, the assets elasticity of consumption (γ_{iA}) shows weak negative correlation

to the earnings elasticity of consumption (γ_{iY}) and the earnings persistence (β_{iY}), which suggests that household's consumption smoothing ability against exogenous earnings and assets shocks can be driven by very different mechanisms.

To examine relations between the elasticities and the observable factors, I estimate correlation between the elasticities and the time-averages of Y_{it} and A_{it} in the dataset. I first estimate cross-moments of the elasticities and the time-averages of (Y_{it}, A_{it}) and then plug them into the correlation formula. The results are presented in Table 2. Correlations to initial values (Y_{i1} and A_{i1}) produced similar results.

Parameters	γ_{iY}	γ_{iA}
$(1/T)\sum_{t=1}^{T} Y_{it}$	[0.049, 0.214]	[-0.045, 0.112]
$(1/T)\sum_{t=1}^{T} A_{it}$	[-0.096, 0.084]	[0.041, 0.252]
$(1/T)\sum_{t=1}^{T} h_{it}$	[-1.051, 0.933]	[-1.083, 0.977]

Table 2: Correlations between the elasticities and the observables. Each entry represents the bounds for the correlation between the elasticity in the column and the data in the row. Replacing the time-averages in the row with their initial values (Y_{i1} and A_{i1}) produced similar results.

The elasticities show positive but weak correlations to time-averages of their respective sources of exogenous shock, and they show no correlation to the other sources. It is notable that the earnings elasticity of consumption shows no correlation to the level of assets. In theory, positive correlation is expected since higher level of assets gives a household less motive for precautionary savings. The empirical result of no correlation suggests that the level of assets is not a main factor that determines earnings elasticity.

The method of this apper also allows to estimate pointwise bounds of the CDFs of the elasticities. The moment conditions listed in Example 5 are used for estimation. Table 3 presents the results, which suggest existence of households with very large (> 1) earnings elasticities of consumption. Households with large earnings elasticities will respond dramatically to policies that induce exogenous changes in earnings.

Parameters	Evaluation Points	-1	-0.5	0	0.5	1	1.5	2
γ_{iY}	LB	0.060	0.149	0.390	0.650	0.815	0.899	0.929
	UB	0.093	0.187	0.449	0.705	0.845	0.993	0.997
γ_{iA}	LB	0.000	0.002	0.278	0.784	0.948	0.984	0.994
	UB	0.016	0.065	0.367	0.815	0.998	1.000	1.000

Table 3: Lower bound ("LB") and upper bound ("UB") of the CDFs of the elasticities at evaluation points.

In the next subsection, I calibrate a structural life-cycle model to enhance understanding of these reduced-form results.

7.2 Coefficient heterogeneity and consumption dynamics

This subsection calibrates a structural model to explain the reduced-form results of the previous subsection. The calibration result will show that earnings persistence heterogeneity is essential for a structural model to generate large heterogeneity in the elasticities of consumption. This suggests that earnings persistence heterogeneity should be considered in life-cycle models in order to reflect real-world consumption behavior accurately.

The dynamic model that I calibrate is similar to the ones used in the simulation exercises of Kaplan and Violante (2010) and Blundell, Low, and Preston (2013). The value function of a household at time t is defined by

$$V_t(A_t, Y_t^*) = \max_{C_t, A_{t+1}} \left[\frac{C_t^{1-\gamma}}{1-\gamma} + \beta \mathbb{E}_t(V_{t+1}(A_{t+1}, Y_{t+1}^*)) \right],$$

subject to

$$0 \leq C_t \leq A_t + Z(t, Y_t^*),$$
$$A_{t+1} \geq -m(t, Y_t^*),$$

where $t \in \{t_0, \ldots, T\}$ is age (with $V_{T+1} = 0$), γ is relative risk-aversion, β is discount factor, C_t is consumption, A_t and A_{t+1} are current and next period asset holdings, Y_t^* is the residual earnings, $Z(t, Y_t^*)$ is the gross earnings, and $m(t, Y_t^*)$ is the borrowing limit. The gross earnings function $Z(t, Y_t^*)$ is defined by

$$Z(t, Y_t^*) = \begin{cases} \exp\{\Gamma(t) + Y_t^*\} & \text{if } t \leq H, \\ \tau_s \times \exp\{\Gamma(H) + Y_H^*\} & \text{if } t > H, \end{cases}$$

where H is retirement age, $\Gamma(t)$ is deterministic trend of earnings which is quadratic in t, and $\tau_s \in [0, 1]$ is a parameter that determines social security benefit payments as a proportion of the last working period's earnings.

I define the borrowing limit $m_A(t, Y_t)$ to be

$$m(t, Y_t) = \begin{cases} \tau_b \times Z(t, Y_t) & \text{if } t \leq H, \\ 0 & \text{if } t > H, \end{cases}$$

That is, the borrowing limit of a household is a τ_b-proportion of current period earnings.

The law of motions of (A_t, Y_t^*) are given by

$$Y_{t+1}^* = \alpha + \rho Y_t^* + \varepsilon_t,$$
$$A_{t+1} = q \times (A_t + Z(t, Y_t^*) - C_t),$$

where α and ρ are scalar, ε_t follows discrete distribution with zero mean, and q is the gross interest rate on the asset.

Except for the earnings process parameters, the structural parameters are assumed to be homogeneous across households. Their calibrated values are summarized in Table 4, which are based on Kaplan and Violante (2010). For the earnings process, I consider two different specifications, namely with and without earnings persistence heterogeneity. Without earnings persistence heterogeneity, households take a common value of $\rho = 0.587$ (which are based on the estimates from PSID; recall that PSID data is measured every two years). With earnings persistence heterogeneity, households take different values of $\rho \in \{0.1,\ 0.85,\ 0.9\}$. In both cases, households take different values of $\alpha \in \{0.2(1-\rho),\ 0,\ 0.2(1-\rho)\}$. Note that the α's are scaled by $1-\rho$ in order to ensure the same stationary mean for every household. In addition, the variance of ε_{it} is set to be $0.041(1-\rho^2)$ which matches the estimates of the PSID dataset. The scaling $1-\rho^2$ is applied to ensure the same stationary variance for every household.

Parameter	Value	Description
β	0.95	discount factor
γ	2	relative risk-aversion
t_0	25	beginning age
T	75	termination age
H	60	retirement age
Γ	2nd order polynomial in age	deterministic trend of earnings
τ_s	0.45	social security benefit parameter
τ_b	0.185	borrowing limit parameter
q	1.03	gross interest rate

Table 4: Description of parameters for the dynamic life-cycle model and their calibrated values which are based on Kaplan and Violante (2010).

All households begin with $A_{t_0} = 0$ and Y_{t_0} is set to be the stationary mean. I solve the model by computing the value functions recursively from the termination period. Details about numerical solution method can be found in the Appendix.

For each specification of the earnings process, I simulate 900 households whose earnings processes are uniformly distributed over the grid of (α, ρ). Then I construct a dataset by collecting data for ages $30, 32, \ldots, 54$, which reflects PSID's biennial sampling. Then the reduced-

form estimation is performed for the dataset, the results of which are presented in Table 5.

Parameter	With heterogeneity	Without heterogeneity
$\mathbb{E}(\gamma_{iY})$	[0.342, 0.408]	[0.299, 0.360]
$\mathbb{E}(\gamma_{iA})$	[-0.017, -0.013]	[-0.019, -0.010]
$\mathbb{E}(\beta_{iY})$	[0.166, 0.441]	[0.251, 0.455]
$\mathrm{Var}(\gamma_{iY})$	[0.323, 0.541]	[0.218, 0.313]
$\mathrm{Var}(\gamma_{iA})$	[0.001, 0.003]	[0.002, 0.008]
$\mathrm{Var}(\beta_{iY})$	[0.002, 0.642]	[0, 0.302]
$\mathrm{Corr}(\gamma_{iY}, \gamma_{iA})$	[0.001, 0.045]	[0.010, 0.076]
$\mathrm{Corr}(\gamma_{iY}, \beta_{iY})$	[-1, 1]	[-1, 1]
$\mathrm{Corr}(\gamma_{iA}, \beta_{iY})$	[-1, -0.058]	[-1, -0.103]

Table 5: Estimates of the parameters of interest based on the datasets simulated from the structural models. The bound estimates are obtained by solving (27) and the corresponding upper bound problem with inflated δ^*.

Table 5 suggests that earnings persistence heterogeneity is essential to yield a large variance estimate of earnings elasticity. Without earnings persistence heterogeneity, the earnings elasticity shows small variance.

In addition, even with earnings persistence heterogeneity, the variance estimates are significantly less than the estimates from the PSID dataset, suggesting additional sources of unobserved heterogeneity. In particular, variance estimate of assets elasticity from the simulated dataset is almost zero even with earnings persistence heterogeneity. This makes sense, since the estimates from the PSID dataset showed weak correlation between the assets elasticity and the earnings persistence. This suggests that an asset-specific source of heterogeneity, such as heterogeneity in interest rates or discount rates, is required to generate heterogeneity in the assets elasticity. Investigating these additional sources of heterogeneity is a topic of future research.

8 Conclusion

This paper studied identification and estimation of dynamic random coefficient models. The model is shown to be not point-identified, and I characterized a sharp identified set using the duality theorem of infinite-dimensional linear programming. A computationally feasible estimation procedure for the identified set is proposed, which exploits linear structure and uses fast and exact algorithm for global polynomial optimization, namely the semidefinite relaxations approach. It is shown that the estimator for the identified set is consistent and that

one can make inference about the identified set using results from the literature on moment inequalities models.

I estimate heterogeneity in earnings persistence and the earnings and asset elasticities of consumption across U.S. households using the PSID dataset. I find that there is large heterogeneity in earnings persistence and the elasticities, and I also find that the earnings persistence and the earnings elasticity of consumption are strongly correlated. However, they show weak correlation to the assets elasticity of consumption. To enhance understanding of these results, I calibrate a structural life-cycle model and estimate reduced-form estimates from the simulated data. It is shown that earnings persistence heterogeneity is essential to generate large heterogeneity observed in the PSID dataset. However, earnings persistence heterogeneity alone is not enough to generate all the features of the PSID estimates, which suggests additional source of unobserved heterogeneity such as interest rate or discount rate. Investigation of such additional sources of heterogeneity is a topic of future research.

Appendices

A Proofs

A.1 Proof of Proposition 1

For simplicity of notation, let's assume that $\mathcal{C} = \mathcal{C}_0^5$ where $\mathcal{C}_0$ is a compact subset of $\mathbb{R}$. The proof can be easily modified for a general 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 to

$$\begin{aligned} \max_{f,g_1,g_2} \ & \mathbb{E}(f(Y_{i0}, Y_{i1}, Y_{i2})) \qquad \text{subject to} \\ & f(Y_{i0}, Y_{i1}, Y_{i2}) + g_1(\gamma_i, \beta_i, Y_{i0})\varepsilon_{i1} + g_2(\gamma_i, \beta_i, Y_{i0}, Y_{i1})\varepsilon_{i2} \leq \beta_i \end{aligned} \tag{33}$$

and the sharp upper bound of $\mathbb{E}(\beta_i)$ equals to

$$\begin{aligned} \min_{f,g_1,g_2} \ & \mathbb{E}(f(Y_{i0}, Y_{i1}, Y_{i2})) \qquad \text{subject to} \\ & f(Y_{i0}, Y_{i1}, Y_{i2}) + g_1(\gamma_i, \beta_i, Y_{i0})\varepsilon_{i1} + g_2(\gamma_i, \beta_i, Y_{i0}, Y_{i1})\varepsilon_{i2} \geq \beta_i. \end{aligned} \tag{34}$$

Now suppose that $\mathbb{E}(\beta_i)$ is point-identified and derive a contradiction. The argument relies on the following proposition.

Proposition 8. *Suppose that $\mathbb{E}(\beta_i)$ is point-identified. Then there exists (f^*, g_1^*, g_2^*) such that*

$$f^*(Y_{i0}, Y_{i1}, Y_{i2}) + g_1^*(\gamma_i, \beta_i, Y_{i0})\varepsilon_{i1} + g_2^*(\gamma_i, \beta_i, Y_{i0}, Y_{i1})\varepsilon_{i2} = \beta_i \tag{35}$$

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

Proof. Suppose such functions do not exist. Then the solution to (33), which is denoted by (f^l, g_1^l, g_2^l), satisfy

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

with positive Lebesgue measure on $\mathcal{C}_0^5$. Similarly, the solution to (34), which is denoted by (f^u, g_1^u, g_2^u), satisfy

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

with positive Lebesgue measure on $\mathcal{C}_0^5$. At least one inequality is strict by assumption. Then it follows that

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

where at least one inequality is strict since the density of $(\gamma_i, \beta_i, Y_{i0}, Y_{i1}, Y_{i2})$ has a lower bound $b > 0$. This implies that 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. □

Now, substitute $\varepsilon_{it} = Y_{it} - \gamma_i - \beta_i Y_{i,t-1}$ in (35) and obtain

$$f^*(Y_{i0}, Y_{i1}, Y_{i2}) + g_1^*(\gamma_i, \beta_i, Y_{i0})(Y_{i1} - \gamma_i - \beta_i Y_{i0}) + g_2^*(\gamma_i, \beta_i, Y_{i0}, Y_{i1})(Y_{i2} - \gamma_i - \beta_i Y_{i1}) = \beta_i. \tag{36}$$

Take any $\gamma, \tilde{\gamma}, \beta, y_0, y_1, y_2 \in \mathcal{C}_0$ such that $\gamma \neq \tilde{\gamma}$. Evaluating (36) at $(\gamma, \beta, y_0, y_1, y_2)$ and

$(\tilde{\gamma}, \beta, y_0, y_1, y_2)$ and taking difference yields

$$\begin{aligned}&(y_1 - \tilde{\gamma} - \beta y_0)\triangle_{\tilde{\gamma},\gamma} g_1^* - (\tilde{\gamma} - \gamma) g_1^*(\gamma, \beta, y_0) \\ &+ (y_2 - \tilde{\gamma} - \beta y_1)\triangle_{\tilde{\gamma},\gamma} g_2^* - (\tilde{\gamma} - \gamma) g_2^*(\gamma, \beta, y_0, y_1) = 0\end{aligned} \tag{37}$$

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

Note that y_2 only appears in the third term of (37). Since (37) must hold almost surely for all $\gamma, \tilde{\gamma}, \beta, y_0, y_1, y_2 \in \mathcal{C}_0$ such that $\gamma \neq \tilde{\gamma}$, it must be that

$$\triangle_{\tilde{\gamma},\gamma} g_2^* = 0 \tag{38}$$

almost surely. 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 (37) with positive measure.

Now (38) implies that g_2^* is almost surely a constant function over γ, i.e.

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

Next, take any $\gamma, \beta, \tilde{\beta}, y_0, y_1, y_2 \in \mathcal{C}$ such that $\beta \neq \tilde{\beta}$, evaluate (36) at $(\gamma, \beta, y_0, y_1, y_2)$ and $(\gamma, \tilde{\beta}, y_0, y_1, y_2)$ and take difference. Then we obtain

$$\begin{aligned}&(y_1 - \gamma - \tilde{\beta} y_0)\triangle_{\tilde{\beta},\beta} g_1^* - (\tilde{\beta} - \beta) y_0 g_1^*(\gamma, \beta, y_0) \\ &+ (y_2 - \gamma - \tilde{\beta} y_1)\triangle_{\tilde{\beta},\beta} g_2^* - (\tilde{\beta} - \beta) y_1 g_2^*(\gamma, \beta, y_0, y_1) = \tilde{\beta} - \beta\end{aligned} \tag{39}$$

where $\triangle_{\tilde{\beta},\beta} g_1^* = g_1^*(\gamma, \tilde{\beta}, y_0) - g_1^*(\gamma, \beta, y_0)$ and $\triangle_{\tilde{\beta},\beta} g_2^*$ is defined similarly. Again, y_2 only appears in the third term, and it follows that

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

Now (37) simplifies to

$$(y_1 - \tilde{\gamma} - \beta y_0)\triangle_{\tilde{\gamma},\gamma} g_1^* - (\tilde{\gamma} - \gamma) g_1^*(\gamma, \beta, y_0) - (\tilde{\gamma} - \gamma) g_2^*(y_0, y_1) = 0. \tag{40}$$

Let $\hat{\gamma} \in \mathcal{C}$ such that $\hat{\gamma} - \tilde{\gamma} = \tilde{\gamma} - \gamma$. Evaluating (40) at $(\gamma, \tilde{\gamma}, \beta, y_0, y_1)$ and $(\tilde{\gamma}, \hat{\gamma}, \beta, y_0, y_1)$ and taking difference yield

$$(y_1 - \hat{\gamma} - \beta y_0)\left(\triangle_{\hat{\gamma},\tilde{\gamma}} g_1^* - \triangle_{\tilde{\gamma},\gamma} g_1^*\right) - (\hat{\gamma} - \tilde{\gamma})\triangle_{\tilde{\gamma},\gamma} g_1^* - (\tilde{\gamma} - \gamma)\triangle_{\tilde{\gamma},\gamma} g_1^* = 0. \tag{41}$$

Since y_1 only appears in the first term, we conclude that

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

almost surely. Then (41) 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 that

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

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

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

Applying similar argument to (39) yields

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

Then (36) 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) in which their coefficients must coincide. In other words, it must be that

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

If we solve this for (g_1^*, g_2^*), we obtain

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

However, g_1^* cannot be a function of y_1, which is contradiction. $\square$

A.2 Proof of Theorem 1

This is a special case of Theorem 2 and hence an immediate consequence of it. Note that, as discussed in Section 5.1.1, it suffices to consider $\lambda > 0$ for the upper bound and $\lambda < 0$ for the lower bound. $\square$

A.3 Proof of Proposition 2

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

In what follows, we 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 given by

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

where μ has the same dimension as R_{it} and λ is scalar. To simplify notation, we abuse notation and let $\mathcal{R}_i = \sum_{t=1}^{T} R_{it}R_{it}'$ and $\mathcal{Y}_i = \sum_{t=1}^{T} R_{it}Y_{it}$, which is without the $1/T$ scaling. The scaling will be applied at the end of the proof.

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

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

Note that the objective function of the inner maximization problem is a quadratic polynomial in v. As discussed in Section 5.1.1, it suffices to consider $\lambda > 0$, in which case the inner maximization problem has a closed-form solution and $\tilde{U}$ simplifies to

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

Expanding the terms gives

$$\begin{aligned} \tilde{U} = \min_{\lambda>0,\, \mu} \Big[& \mu'\mathbb{E}(\mathcal{Y}_i) + \frac{1}{4\lambda} e'\mathbb{E}(\mathcal{R}_i^{-1})e + \frac{\lambda}{4}\mathbb{E}(\mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i) + \frac{1}{4\lambda}\mu'\mathbb{E}(\mathcal{R}_i)\mu \\ & + \frac{1}{2}e'\mathbb{E}(\mathcal{R}_i^{-1}\mathcal{Y}_i) - \frac{1}{2\lambda}e'\mu - \frac{1}{2}\mu'\mathbb{E}(\mathcal{Y}_i) \Big]. \end{aligned} \tag{42}$$

We first 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

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

Substitutes this into (42) yields

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

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 given by

$$\lambda = \sqrt{\frac{e'\mathbb{E}(\mathcal{R}_i^{-1})e - e'\mathbb{E}(\mathcal{R}_i)^{-1}e}{\mathbb{E}(\mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i) - \mathbb{E}(\mathcal{Y}_i)'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i)}}. \tag{43}$$

The numerator and the denominator inside the square root are all weakly positive. In addition, 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, for matrices R and Y that have the same dimensions as $\mathcal{R}_i$ and $\mathcal{Y}_i$, define the functions

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

Similar to the scalar case where the inverse function $f(x) = 1/x$ is convex, one can show that these functions are convex. In particular, the following result is known[9]:

Lemma 1 (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*

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

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*

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

Then we can apply Jensen's inequality to $\mathcal{E}(R)$ and $\mathcal{D}(Y, R)$ and show that the numerator and the denominator in (43) are weakly positive.

This finishes derivation of optimal μ and λ of (42). Then we can plug in the optimal μ and λ into (42) and scale the terms by $1/T$ to obtain the expression for $\tilde{U}$ in Proposition 2. □

A.4 Proof of Proposition 3

The proof focuses on proving that $L_M \to -\infty$ as $M \to \infty$. The argument can be applied symmetrically to U_M.

For a given M, let $g_t(Z_i, X_i^t)$ be a function from $[-M, M]^{qT+pt}$ to $\mathbb{R}$. By Theorem 3, L_M is given by

$$\begin{aligned} L_M &= \max_{g_1,\ldots,g_T} \mathbb{E}\left(\min_v \left[e'v + \sum_{t=1}^{T} g_t(Z_i, X_i^t)(Y_{it} - R_{it}'v)\right]\right) \\ &= \max_{g_1,\ldots,g_T} \left\{\mathbb{E}\left(\sum_{t=1}^{T} g_t(Z_i, X_i^t) Y_{it}\right) + \mathbb{E}\left(\min_v \left[e - \sum_{t=1}^{T} g_t(Z_i, X_i^t) R_{it}\right]' v\right)\right\} \end{aligned}$$

where the expectation is with respect to P_M. One can show that

$$\min_v \left[e - \sum_{t=1}^{T} g_t(Z_i, X_i^t) R_{it}\right]' v \leq -M\mathbf{1}' \left|e - \sum_{t=1}^{T} g_t(Z_i, X_i^t) R_{it}\right|,$$

where the absolute value operator $|\cdot|$ is applied element-wise and $\mathbf{1}$ is the vector of ones. The reason why is that the minimum of a function is less than a function value evaluated at a point

$$v = -M \times \operatorname{sgn}\left(e - \sum_{t=1}^{T} g_t(Z_i, X_i^t) R_{it}\right)$$

where $\text{sgn}(\cdot)$ is a sign function applied element-wise to a vector. Then it follows that

$$L_M \leq \tilde{L}_M \equiv \max_{g_1,\cdots,g_T} \left\{ \mathbb{E}\left(\sum_{t=1}^{T} g_t(Z_i, X_i^t) Y_{it} \right) - M\mathbb{E}\left(\mathbf{1}' \left| e - \sum_{t=1}^{T} g_t(Z_i, X_i^t) R_{it} \right| \right) \right\}.$$

Note that the argument (Z_i, X_i^t) of g_t contains $R_{it} = (Z_{it}, X_{it})$, which means that g_t can be chosen to be R_{it}-specific. Note also that, when M is large, the choice of $\{g_t\}$ affects $\tilde{L}_M$ mainly through the second term of $\tilde{L}_M$. This means that the optimal $\{g_t\}$, denoted by $\{g_t^*\}$, for each given $(R_{i1}, \ldots, R_{iT})$ is given approximately by (for large M)

$$(g_1^*, \ldots, g_T^*) \approx \underset{g_1,\cdots,g_T}{\text{argmin}} ||e - \sum_{t=1}^{T} g_t(Z_i, X_i^t) R_{it}|| = \underset{g}{\text{argmin}} ||e - \mathbf{R}_i g||_2$$

where $||\cdot||_2$ is the L^2 norm, $g = (g_1, \ldots, g_T)$ and $\mathbf{R_i} = (R_{i1}, \ldots, R_{iT})$. Therefore, $g^* = (g_1^*, \ldots, g_T^*)$ is given approximately by (for large M)

$$g^* \approx \left(\mathbf{R}_i \mathbf{R}_i'\right)^{-1} \mathbf{R}_i e.$$

Then we have, for large M,

$$\tilde{L}_M \approx \mathbb{E}\left(\sum_{t=1}^{T} e' R_{it} Y_{it} \right) - M \times \mathbf{1}' \mathbb{E}(|(I - \mathbf{R}_i' \left(\mathbf{R}_i \mathbf{R}_i'\right)^{-1} \mathbf{R}_i) e|). \tag{44}$$

where I is an identity matrix. Now assume the following regularity conditions:

Assumption 6. The sequence of distributions $\{P_M\}$ satisfies the following conditions.

- $\mathbb{E}\left(\sum_{t=1}^{T} e' R_{it} Y_{it}\right)$ converges to a finite number as $M \to \infty$.
- $\mathbb{E}(|(I - \mathbf{R}_i' \left(\mathbf{R}_i \mathbf{R}_i'\right)^{-1} \mathbf{R}_i) e|)$ converges to a nonzero finite vector as $M \to \infty$.

Under Assumption 6, the right-hand side of (44) tends to $-\infty$ as $M \to \infty$, which implies that $L_M \to -\infty$. □

A.5 Proof of Theorem 2

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

In summary, (10) can be obtained by taking the dual of (9). The main part of this proof is to show that the conditions for the duality theorem hold, which requires additional regular-

ity conditions. This is different from finite-dimensional linear programming (LP) where the regularity conditions for the duality theorem always hold.

Assume the following regularity conditions.

Assumption 7. 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 compactness of $\mathcal{W} \times \mathcal{V}$. The third condition means that the data generating process of the model, or its observationally equivalent one, has full support. This condition is called the interior point condition required for the duality to hold in an infinite-dimensional LP. Note that the interior point condition implies that (9) is feasible and that its solution is finite.

In what follows, we show that a key condition for the duality of infinite-dimensional LP holds, for which we 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 $\overline{\mathcal{F}}_{W \times V}$ since 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 $\overline{\mathcal{F}}_W$ to be the dual space of $\mathcal{M}_W$ and define $\mathcal{F}_W$ to 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 $h = (h_1, \ldots, h_K, f_h)$ to denote their generic elements. Note that $\mathcal{H}$ is a dual space of $\mathcal{G}$. Define the dual pairing

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

Now define a linear map $A : \mathcal{M}_{W\times V} \mapsto \mathcal{G}$ to be

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

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 that

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

Then it follows that

$$\langle A(P), h \rangle = \sum_{k=1}^{K} h_k \int \phi_k dP + \int f_h dP = \int \left[\sum_{k=1}^{K} h_k \phi_k + f_h \right] dP \equiv \langle P, A^*(h) \rangle, \tag{45}$$

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 (45) shows that A is weakly continuous with adjoint A^*, which is a key condition for the duality to hold.

Now we rewrite (9) into the standard form of infinite-dimensional LP:

$$\min_{P\in\mathcal{M}_{W\times V}} \langle P, m \rangle \qquad \text{subject to} \qquad A(P) = c, \qquad P \geq 0, \tag{46}$$

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

Now, with Assumption 7 and A being weakly continuous, the strong duality holds for (46)[10], i.e. the optimal solution to (46) equals to the solution to the following problem:

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

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

which can be written more concretely as

$$\max_{h_1,\ldots,h_K\in\mathbb{R},\ f_h\in\overline{\mathcal{F}}_W} \int f_h dP_W \qquad \text{subject to} \qquad \sum_{k=1}^{K} h_k\phi_k + f_h \leq m. \tag{47}$$

We can be solve for f_h to simplify (47). Rearrange the constraint of (47) and obtain

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

Since the left-hand side is only a function of w, it follows that

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

on $\mathcal{W}$. Then, since the objective of (47) is to maximize the integral of $f_h(w)$, it must be that the solution f_h^* satisfies

$$f_h^*(w) = \min_{v\in\mathcal{V}} \left[m(w,v) - \sum_{k=1}^{K} h_k\phi_k(w,v) \right] \tag{48}$$

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 we can increase the value of f_h^* by an infinitesimal amount and increase the value of the objective, which is contradiction.

Substituting (48) into (47) yields the program

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

Since $(h_1,\ldots,h_K)$ is a choice variable supported on $\mathbb{R}^K$, the problem remains equivalent even if we switch the signs of $(h_1,\ldots,h_K)$. If we 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.6 Proof of Proposition 4

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

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

which is the definition of concavity. □

A.7 Proof of Proposition 5

By the proof of Proposition 4, $\hat{L}(\lambda)$ and $L(\lambda)$ are concave. Then, as in the proof of Theorem 2.7 in Newey and McFadden (1994), uniform convergence of $\hat{L}$ to L holds on any compact set $K \subseteq \mathbb{R}^K$:

$$
\sup_{\lambda \in K} |\hat{L}(\lambda) - L(\lambda)| \xrightarrow{p} 0. \tag{49}
$$

Let $\hat{\lambda} = \operatorname{argmax}_\lambda \hat{L}(\lambda)$ and $\lambda_0 = \operatorname{argmax}_\lambda L(\lambda)$. If there are multiple argmax's, choose any of them. Then it follows that

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

Then, 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 (49).

Now, let Λ_0 be the set of all $\operatorname{argmax}_\lambda L(\lambda)$. 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 and hence the consistency is immediate. If such ε exists, then 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 the Euclidean distance. This implies that $\hat{\lambda} \in K_0$ with probability approaching to one. □

A.8 Proof of Proposition 6

One can rewrite (25) 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 one can replicate the argument of Theorem 2 and show that (26) follows by taking the dual of the above and simplifying it. □

A.9 Proof of Proposition 7

One can rewrite (28) as

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

Then one can replicate the argument of Theorem 2 and show that (27) follows by taking the dual of the above and simplifying it. □

B Generalization of Theorem 2 with conditional moment restrictions

This subsection extends the model in Assumption 5 and also consider conditional moment restrictions. Concretely, consider the following model.

Assumption 8. The random vectors (W_i, V_i) satisfy

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

where the ϕ_k's and the 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.

Compared to Assumption 5, there are conditional moment restrictions represented by (ψ_k, A_k). The objective of this subsection is to obtain the counterpart of Theorem 2 under Assumption 8.

Consider the problem of characterizing the identified set I of θ which is defined by

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

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 coordinates of (W_i, V_i). Let $\mathcal{A}_k$ be the support of A_{ik}, which is the projection of $\mathcal{W} \times \mathcal{V}$ onto the coordinates of A_{ik}, and let $\mathcal{A}'_k$ be the support of A'_{ik}. Then (A_{ik}, A'_{ik}) partitions (W_i, V_i) into two sets of coordinates. We abuse notation and write any function $f(w, v)$ on $\mathcal{W} \times \mathcal{V}$ also as $f(a_k, a'_k)$ on $\mathcal{A}_k \times \mathcal{A}'_k$ for any k.

Now we have the following result.

Theorem 3. *Suppose Assumption 8 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 $\mu_k : \mathcal{A}_k \mapsto \mathbb{R}$,*

$$L = \max_{\{\lambda_k\}_{k=1}^{K_U}, \{\mu_k\}_{k=1}^{K_C}} \mathbb{E}\left[\min_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K_U} \lambda_k \phi_k(W_i, v) + \sum_{k=1}^{K_C} \mu_k(A_k(W_i, v))\psi_k(W_i, v) \right\}\right] \quad (50)$$

and

$$U = \min_{\{\lambda_k\}_{k=1}^{K_U}, \{\mu_k\}_{k=1}^{K_C}} \mathbb{E}\left[\max_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K_U} \lambda_k \phi_k(W_i, v) + \sum_{k=1}^{K_C} \mu_k(A_k(W_i, v))\psi_k(W_i, v) \right\}\right] \quad (51)$$

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

Proof. The proof focuses on (50). The same argument applies to the upper bound. Also, we abuse notation and 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, in the third equation, we represent the conditional moment restriction using integral over $\mathcal{A}_k'$.

The lower bound of I is given by the program

$$\begin{aligned} \min_{p \in \mathcal{M}_{W\times V},\, p \geq 0} & \int m(w,v)p(w,v)d(w,v) \qquad \text{subject to} \\ & \int \phi_k(w,v)p(w,v)d(w,v) = 0, \quad k = 1, \ldots, K_U, \\ & \int \psi_k(a_k, a_k')p(a_k, a_k')da_k' = 0, \text{ for all } a_k \in \mathcal{A}_k, \quad k = 1, \ldots, K_C, \\ & \int p(w,v)dv = p_W(w) \;\text{ for all } w \in \mathcal{W}. \end{aligned} \quad (52)$$

In what follows, as in the proof of Theorem 2, we assume additional regularity conditions and check that the conditions for the duality theorem hold. Then (50) is obtained by taking the dual of (52).

Assume the following regularity conditions.

Assumption 9. 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 (52) 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 9) are the counterparts of Assumption 7. Note that, since $\mathcal{W} \times \mathcal{V}$ is compact, the second condition implies that they are L^p with respect to the Lebesgue measure for any $p \geq 1$. The fourth condition is restrictive. However, we use Theorem 3 to prove non-identification results of Propositions 1 and 3, and for this purpose it is enough to show Theorem 3 under this condition.

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

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

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

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 that

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

and

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

Then it follows that

$$\langle A(P), h \rangle = \int \left[\sum_{k=1}^{K_U} h_k \phi_k + \sum_{k=1}^{K_C} \overline{h}_k \psi_k + f_h \right] p(w,v) d(w,v). \equiv \langle p, A^*(h) \rangle, \tag{53}$$

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 (53) shows that A is weakly continuous with adjoint 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 9 and the weak continuity of A, and the optimal solution to (52) equals to the solution to the following dual problem:

$$\max_{h_1,\ldots,h_{K_U},\overline{h}_1,\ldots,\overline{h}_{K_C},f_h} \int f_h(w) p_w(w) dw \qquad \text{subject to} \qquad \sum_{k=1}^{K_U} h_k \phi_k + \sum_{k=1}^{K_C} \overline{h}_k \psi_k + f_h \leq m. \tag{54}$$

Then, 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 (50).

□

C Global polynomial optimization

This subsection formally discuss the theory of global polynomial optimization discussed in Section 5.1.2. 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 $v = (v_1, \ldots, v_m) \in \mathbb{R}^m$ and 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:

$$\{v_1^{\alpha_1} v_2^{\alpha_2} \cdots v_m^{\alpha_m} \mid \alpha_1, \ldots, \alpha_m \in \mathbb{N}\}$$

where $\mathbb{N} = \{0, 1, 2, \ldots\}$. Let $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{N}^m$ and $v^\alpha = v_1^{\alpha_1} v_2^{\alpha_2} \cdots v_m^{\alpha_m}$. Define $|\alpha| = \alpha_1 + \cdots + \alpha_m$ be the degree of v^α.

A polynomial $f(v)$ in v can be written in the standard form

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

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 $d < \infty$. We are interested in minimizing f with respect to v:

$$\min_{v \in \mathbb{K}} f(v) = \min_{v \in \mathbb{K}} \sum_{\alpha \,:\, |\alpha| \leq d} c_\alpha v^\alpha, \tag{55}$$

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

$$\mathbb{K} = \{v \in \mathbb{R}^m \mid g_j(v) \geq 0,\ j = 1, \ldots, J\} \tag{56}$$

where g_j's are polynomials in v. Let $d_j < \infty$ be the degree of g_j. If $\mathbb{K}$ is given as in (56), we assume that $\mathbb{K}$ is compact.

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[11] transforms (55) into a convex optimization problem. Let $\tilde{\mathcal{V}}$ be the space of Borel measures in $\mathbb{R}^m$ whose supports are contained in $\mathbb{K}$. We can write

$$\min_{v \in \mathbb{K}} f(v) = \min_{P \in \tilde{\mathcal{V}},\ \int dP = 1} \int f(v) dP = \min_{P \in \tilde{\mathcal{V}},\ \int dP = 1} \sum_{\alpha \,:\, |\alpha| \leq d} c_\alpha \int v^\alpha dP \tag{57}$$

where the optimal P is the point-mass distribution concentrated at the minimizer of $f(v)$. Let $\{y_\alpha\}$ be a sequence of real numbers indexed by α. We can rewrite (57) as

$$\min_{\{y_\alpha\}} \sum_\alpha c_\alpha y_\alpha \qquad \text{subject to} \qquad y_\alpha = \int v^\alpha dP \quad \text{for some } P \in \tilde{\mathcal{V}},\ \int dP = 1. \tag{58}$$

The objective function in (58) is linear in $\{y_\alpha\}$, which means that it is a convex objective function in $\{y_\alpha\}$. It remains to characterize the constraint of (58) as a convex constraint in $\{y_\alpha\}$.

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

C.2 The moment problem

The constraint in (58) is very closely related to the problem called the *moment problem* in mathematics. It 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 $\alpha \geq 0$?" 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 result is known.

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*

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

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

The Riesz-Haviland theorem provides a characterization of the constraint in (58) which only involves $\mathbf{y} = \{y_\alpha\}_{\alpha\in\mathbb{N}^m}$. However, checking $L_\mathbf{y}(f) \geq 0$ for all nonnegative f is computationally infeasible. This motivates us to find a tractable characterization that can be derived 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 given by

$$u_r(v) = (1, v_1, \ldots, v_m, v_1^2, v_1v_2, \ldots, v_m^2, \ldots, v_1^r, \ldots, v_m^r)'.$$

Let s_r be the length of $u_r(v)$, which equals to $(n+r)$-choose-n, i.e. $(n+r)!/n!r!$. Then a polynomial in v whose degree is at most r can be represented by a 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 $u_r(v)u_r(v)'$ involves polynomials of degree at most $2r$. Alternatively, 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

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

Let $p, q \in \mathcal{V}_r$ with coefficient vectors $\mathbf{p}$ and $\mathbf{q}$. The following result holds:

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

Suppose $\mathbf{y}$ has a representing measure. Then it follows that

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

by the Riesz-Haviland theorem, which implies that $M_r(\mathbf{y})$ 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

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

Alternatively, 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 given by

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

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

$g_j \geq 0$ on $\mathbb{K}$, then it follows that

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

by the 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 (56) and is compact. Note that, if $\mathbb{K} \in \mathbb{R}^m$ is compact, there exists a real number $B > 0$ such that

$$B - ||v||^2 = B - \sum_{k=1}^{m} v_k^2 \geq 0$$

on $\mathbb{K}$. The following theorem provides a necessary and sufficient condition for the moment problem.

Theorem 5 (Putinar's Positivstellensatz)**.** *Let $\mathbb{K}$ be defined as in (56) 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 (56) and is compact, the polynomial optimization problem in (58) is equivalent to the following problem:

$$\begin{aligned} V = \min_{\mathbf{y}} \sum_{\alpha} c_\alpha y_\alpha \qquad \text{subject to} \qquad & y_0 = 1, \\ & M_r(\mathbf{y}) \text{ is PSD for all } r, \\ & M_r(g_j \mathbf{y}) \text{ is PSD for } j = 1, \ldots, J+1, \text{ for all } r. \end{aligned} \tag{59}$$

It turns out that the PSD constraints are convex constraints, which is based on the fact that the set of semidefinite matrices is a convex set (see Lemma 1 in the proof of Proposition 2). Then (59) is an optimization problem of a linear objective subject to convex constraints. Moreover, the constraints belong to a special class of convex constraints so that (59) has its own name: semidefinite program (SDP).

Unfortunately, (59) has infinite number of constraints, which means that 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

$$
\begin{aligned}
V_r = \min_{\mathbf{y}_{2r}} \sum_{|\alpha| \leq 2r} c_\alpha y_\alpha \qquad \text{subject to} \qquad & y_0 = 1, \\
& M_r(\mathbf{y}_{2r}) \text{ is PSD}, \\
& M_{r-r_j}(g_j \mathbf{y}_{2r}) \text{ is PSD for all } j = 1, \ldots, J+1.
\end{aligned}
\tag{60}
$$

Note that $M_r(\mathbf{y}_{2r})$ being PSD implies $M_{r'}(\mathbf{y}_{2r})$ being PSD for all $r' \leq r$. We can solve (60) using SDP solvers available in the industry.

The constraint in (60) is a finite subset of the constraint in (59). This means that $V_r \leq V$, i.e. V_r is a lower bound for the global minimum of the polynomial. Also, V_r is monotonically increasing in r since there are more constraints as r increases, and we can show 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. We now introduce a method for checking it, which is called the *certificate* of optimality. The method is closely related to 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 $\mathbf{y}_{2r}$ admits a $\operatorname{rank}(M_r(\mathbf{y}_{2r}))$-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,*
- *$\operatorname{rank}(M_r(\mathbf{y}_{2r})) = \operatorname{rank}(M_{r-\bar{r}}(\mathbf{y}_{2r}))$.*

With this theorem, we have the following algorithm for solving constrained polynomial optimization problem.

1. Set $r = \lfloor (d+1)/2 \rfloor$.

2. Solve (60) and compute V_r and $\mathbf{y}_{2r}$.

3. Check if $\text{rank}(M_r(\mathbf{y}_{2r})) = \text{rank}(M_{r-\bar{r}}(\mathbf{y}_{2r}))$.

4. If [3.] is true, then V_r is the exact minimum of the polynomial where the number of minimizers equals to $\text{rank}(M_r(\mathbf{y}_{2r}))$. If [3.] is false, increase r by 1 and go to [2.].

When implementing this algorithm, 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. Consider the following program:

$$V_r^* = \min_{\mathbf{y}_{2r}} \sum_{|\alpha| \leq 2r} c_\alpha y_\alpha \qquad \text{subject to} \qquad \begin{aligned} & y_0 = 1, \\ & M_r(\mathbf{y}_{2r}) \text{ is PSD.} \end{aligned} \tag{61}$$

As discussed previously, the constraint in (61) 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 $f = \sum_{|\alpha| \leq 2r} c_\alpha y_\alpha$. Also, the following theorem states certificate of optimality for unconstrained polynomial optimization problem.

Theorem 7 (Lasserre, 2010, Theorem 3.7)**.** *Let $\mathbf{y}_{2r} = \{y_\alpha\}_{\alpha \in \mathbb{N}^m_{2r}}$. Then the sequence $\mathbf{y}_{2r}$ admits 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 $M_{r+1}(\mathbf{y}_{2r+2})$.*

Now 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 (61) 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 where 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 where 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 certificate of optimality based on Lasserre (2015, Theorem 2.36), which is easier to check than the flat extension condition in Step [4.].

Note that the algorithm for the unconstrained case solves only one SDP which do not guarantee the exact solution. Nie, Demmel, and Sturmfels (2006) proposed a refinement of the algorithm where we 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 hence the first order condition

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

is a polynomial constraint. In addition, for any polynomial $g(v)$, the first order condition generalizes to

$$g(v)\frac{\partial f}{v_k} = 0, \quad k = 1, \ldots, m, \tag{62}$$

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 (62) as an additional constraint for (61) for all monomials $g(v)$ of degree at most $r - \lfloor (d+1)/2 \rfloor$. 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 $\mathbf{y}_{2r}$. My R package optpoly implements the algorithm for extraction of the solutions. The Gloptipoly package for Matlab also has an implementation of the algorithm. For the details about the algorithm, refer to Lasserre (2010, 2015).

The solution extraction algorithm requires knowledge of $\text{rank}(M_r(\mathbf{y}_{2r}))$, which equals to the number of minimizers. The rank of this matrix can be checked by numerically counting zero eigenvalues, but 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. In this case we can assume $\text{rank}(M_r(\mathbf{y}_{2r})) = 1$ and execute the algorithm.

When $\text{rank}(M_r(\mathbf{y}_{2r})) = 1$, i.e. when the polynomial has a unique minimizer, the extraction algorithm becomes very simple. It extracts the minimizer $v^* = (v_1^*, \ldots, v_m^*) = \text{argmin}_v f(v)$ by the following rule:

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

where $\mathbf{e}_k$ is the vector with one at the k-th entry and zero elsewhere. The idea is the following. If $\text{rank}(M_r(\mathbf{y}_{2r})) = 1$, then $\mathbf{y}_{2r}$ has a 1-atomic representing probability measure P^*, which 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 Computation of the critical value

This subsection describes computation of the critical value used for the inference procedure described in Section 6.

To compute the critical value, draw bootstrap samples $(W_1^{(b)}, \ldots, W_N^{(b)})$ of $(W_1, \ldots, W_N)$ for $b = 1, \ldots, B$, and compute

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

and

$$\hat{S}^{(b)}(\lambda) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(G(\lambda, W_i^{(b)}) - \hat{L}^{(b)}(\lambda) \right)^2}$$

for each bootstrap sample. Then define

$$c_\xi^{(1)} = \ (1-\xi)\text{-quantile of } \left\{ \max_{\lambda \in \Lambda_F} \frac{\sqrt{N}(\hat{L}(\lambda) - \hat{L}^{(b)}(\lambda))}{\hat{S}^{(b)}(\lambda)} \right\}_{b=1}^{B}.$$

Next, compute for each $\lambda \in \Lambda_F$

$$u_\lambda = \hat{L}(\lambda) + \hat{S}(\lambda) \times \frac{c_\xi^{(1)}}{\sqrt{N}},$$

Then the critical value $c^*(\alpha, \xi)$ is given by

$$c^*(\alpha, \xi) = \ (1 - \alpha + \xi)\text{-quantile of } \left\{ \max_{\lambda \in \Lambda_F} \frac{\sqrt{N}(\hat{S}^{(b)}(\lambda) - \hat{L}(\lambda) + u_\lambda)}{\hat{S}^{(b)}(\lambda)} \right\}_{b=1}^{B}.$$

E Construction of the dataset

The raw data is obtained from Panel Data of Income Dynamics (PSID). In particular, I use PSID's Main Study data from 1999 to 2015, which is sampled every two years. I use the *Family Files* for years 1999-2015, and I also use the *Wealth files* for years 1999-2007 since the wealth data are separately supplied for these years from the PSID website. These files contain information on earnings, consumption and asset holdings of U.S. households. Year-of-birth and education level of the household head is found in the *Cross-year Individual file*, which can be merged to Family Files using the household ID.

In order to avoid oversampling of low-income households, I drop the households from Survey of Economic Opportunity (SEO). Also, I only keep married households with male household head whose age is between 20 and 65.

The price index data is taken from the Consumer Price Index (CPI), which is downloaded from the Federal Reserve Back of St. Louis. The price index of all urban consumers is used.

Then the consumption is defined to be the sum of nondurables and services expenditure and the consumption of food stamps, divided by the price index. The earnings is defined to be sum of the earnings of the married couple divided by the price index. I directly use the *Constructed wealth variable including equity* entry of the PSID dataset and define the asset to be this entry divided by the price index. It is the sum of values of seven asset types, net of their debt values, plus the value of home equity.

What I use in the estimation and the calibration in Section 7 is the natural logs of them net of demographic dummies. Concretely, I compute the residuals of log-consumption, log-earnings and log-assets by regressing them on the dummies of education, year-of-birth, wife's education, wife's year-of-birth, state, family size, number of kids under age 18, race and the indicator of the child support not living with the household head.

Then I drop individuals with missing information on these dummies or missing information on the consumption, the earnings or the assets. This gives a balanced panel of $N = 684$ individuals with 9 waves. Since the lagged earnings is used as a regressor in the reduced-form model, the first wave is used for the lagged earnings and I am left with $T = 8$ waves.

F Sensitivity and robustness checks of the reduced-form estimates

This subsection performs sensitivity and robustness checks of the reduced-form estimates presented in Section 7.1. There are two exercises that are performed. First, I check sensitivity of the estimates against sample variation. Second, I check robustness of the estimates against small within variation.

F.1 Sensitivity check against sample variation

The first exercise concerns sensitivity of the estimates against sample variation, which is decribed in what follows. First, draw bootstrap samples, namely the datasets that have the same size as the PSID dataset which is created by sampling individuals in the PSID dataset with replacement. Then, compute the estimate for each bootstrap sample and summarize the distribution of the estimates.

Table 6 summarizes distribution of 500 bootstrap estimates. The results are presented for mean of earnings persistence $\mathbb{E}(\beta_{iY})$. Other parameters of interest stated in Table 1 showed similar level of sensitivity. One can see that bootstrap sample variation does not vary the estimate significantly.

Parameter: $\mathbb{E}(\beta_{iY})$	Mean	SD	Min	Max	5%	95%
LB	0.341	0.049	0.151	0.467	0.265	0.425
UB	0.369	0.050	0.187	0.504	0.292	0.453

Table 6: Distribution of lower and upper bounds for $\mathbb{E}(\beta_{iY})$ across 500 bootstrap samples. The estimates are obtained by solving (27) and the corresponding upper bound problem with inflated δ^*.

F.2 Robustness check against small within variation

The second exercise concerns robustness of the estimates against small within variation, that is, near-multicollinearity of individual time series. Recall that Assumption 2 for identification of means requires that there is no multicollinearity. The second exercise checks robustness against violation of this assumption.

The exercise is described in what follows. Note that the model can be written in the matrix

form

$$\begin{pmatrix} C_{it} \\ Y_{i,t+1} \end{pmatrix} = \begin{pmatrix} 1 & Y_{it}^* & A_{it} & 0 & 0 \\ 0 & 0 & 0 & 1 & Y_{it} \end{pmatrix} \begin{pmatrix} \gamma_{i0} \\ \gamma_{iY} \\ \gamma_{iA} \\ \beta_{i0} \\ \beta_{iY} \end{pmatrix} + \begin{pmatrix} \nu_{it} \\ \varepsilon_{it} \end{pmatrix}.$$

For each individual, I compute time series average of the design matrix $(1/T)\sum_{t=1}^{T} \mathbf{R}_{it}'\mathbf{R}_{it}$ and compute its minimum eigenvalue. Then I drop 10% of individuals from the PSID dataset with smallest minimum eigenvalues and compute the reduced-form estimates with the remaining individuals.

Table 7 summarizes the result. The estimates under trimming is not very different to estimates under the full PSID dataset in Table 1, which suggests that the estimates are robust against near-multicollinearity of individuals.

Parameter	LB	UB
$\mathbb{E}(\gamma_{iY})$	0.181	0.243
$\mathbb{E}(\gamma_{iA})$	0.140	0.181
$\mathbb{E}(\beta_{iY})$	0.352	0.365
$\mathrm{Var}(\gamma_{iY})$	0.693	0.811
$\mathrm{Var}(\gamma_{iA})$	0.157	0.195
$\mathrm{Var}(\beta_{iY})$	0.458	0.537
$\mathrm{Corr}(\gamma_{iY}, \gamma_{iA})$	-0.123	-0.050
$\mathrm{Corr}(\gamma_{iY}, \beta_{iY})$	0.060	0.471
$\mathrm{Corr}(\gamma_{iA}, \beta_{iY})$	-0.130	0.063

Table 7: Estimates for the parameters of interest, with dataset trimming. The lower bound ("LB") and the upper bound ("UB") are obtained by solving (27) and the corresponding upper bound problem with inflated δ^*.

G Numerical solution method for dynamic life-cycle model

The solution method follows that of Kaplan and Violante (2014). It takes as an input

- A discrete grid of (A_t, Y_t^*), denoted by $(a_j, y_j)_{j=1}^{J}$,
- A discrete grid of ε_{it}, denoted by $(e_s)_{s=1}^{S}$,

and it produces as an output

- A grid of values of $V_t(A_t, Y_t^*)$ on the discrete grid of (A_t, Y_t^*) for each $t = t_0, \ldots, T$.

Given discrete grid $(e_s)_{s=1}^S$ of ε_{it}, the value function is given by

$$V_t(A_t, Y_t) = \max_{C_t, A_{t+1}} \left[\frac{C_t^{1-\gamma}}{1-\gamma} + \beta \frac{1}{S} \sum_{s=1}^{S} V_{t+1}(A_{t+1}, \alpha + \rho Y_t^* + e_s) \right].$$

To describe the method, it is convenient to write the value function equation in terms of state variables only. Concretely, rewrite the value function as

$$V_t(A_t, Y_t^*) = \max_{A_{t+1}} \left[\frac{(A_t + Z(t, Y_t^*) - q^{-1} A_{t+1})^{1-\gamma}}{1-\gamma} + \beta \frac{1}{S} \sum_{s=1}^{S} V_{t+1}(A_{t+1}, \alpha + \rho Y_t^* + e_s) \right], \quad (63)$$

where $C_t = A_t + Z(t, Y_t^*) - q^{-1} A_{t+1}$ is substituted.

Let $B_t(A_t, Y_t^*)$ be the argmax of (63) and $D_t(A_t, Y_t^*) = \partial V_t(A_t, Y_t^*)/\partial A_t$ be the derivative of V_t with respect to A_t. For each point (a, y) in the discrete grid of (A_t, Y_t^*), record

$$\begin{aligned} B_T(a, y) &= 0, \\ V_T(a, y) &= \frac{(a + Z(T, y))^{1-\gamma}}{1-\gamma}, \\ D_T(a, y) &= (a + Z(T, y))^{-\gamma}. \end{aligned}$$

Then, for each $t = T-1, T-2, \ldots, t_0$ given the grid of $(B_{t+1}, V_{t+1}, D_{t+1})$, compute the values of (B_t, V_t, D_t) for each point (a, y) in the discrete grid of (A_t, Y_t^*). The first step is to compute $B_t(a, y)$ as the solution b^* to the Euler equation

$$q^{-1}(a + Z(t, y) - q^{-1} b^*)^{-\gamma} = \beta \frac{1}{S} \sum_{s=1}^{S} D_{t+1}(b^*, \alpha + \rho y + e_s)$$

subject to the constraint $-m(t, y) \leq b^* \leq q^{-1}(a + Z(t, y))$, where the value of $D_{t+1}(\cdot, \cdot)$ is obtained by bilinear interpolation. Concretely, create a grid $\{b_u\}_{u=1}^U$ of values of b where

$$b_u = -m(t, y) \times \frac{U - u}{U} + q^{-1}(a + Z(t, y)) \times \frac{u}{U}.$$

and choose $b^* = b_{u^*}$ where

$$u^* = \operatorname*{argmin}_u \left| q^{-1}(a + Z(t, y) - q^{-1} b_u)^{-\gamma} - \beta \frac{1}{S} \sum_{s=1}^{S} D_{t+1}(b_u, \alpha + \rho y + e_s) \right|.$$

In words, set b^* to be the value of b_u that minimizes difference between the LHS and the RHS of the Euler equation.

Then compute $V_t(a, y)$ by

$$V_t(a,y) = \frac{(a + Z(t, Y_t^*) - q^{-1}B_t(a,y))^{1-\gamma}}{1-\gamma} + \beta \frac{1}{S} \sum_{s=1}^{S} V_{t+1}(B_t(a,y), \alpha + \rho y + e_s)$$

where the values of $B_t(\cdot,\cdot)$ and $V_{t+1}(\cdot,\cdot)$ are obtained by bilinear interpolations. Lastly, compute $D_t(a, y)$ from the Envelope condition

$$D_t(a,y) = (a + Z(t, Y_t^*) - q^{-1}B_t(a,y))^{-\gamma}.$$

This completes computation of (B_t, V_t, D_t) for a given value of t.

Simulating data from the model is performed similarly. Given values of (A_t, Y_t^*, t), create a grid $\{b_u\}_{u=1}^{U}$ of values of b where

$$b_u = -m(t,y) \times \frac{U-u}{U} + q^{-1}(A_t + Z(t, Y_t^*)) \times \frac{u}{U},$$

and choose $A_{t+1} = b_{u^*}$ where

$$u^* = \underset{u}{\operatorname{argmin}} \left| q^{-1}(A_t + Z(t, Y_t) - q^{-1}b_u)^{-\gamma} - \beta \frac{1}{S} \sum_{s=1}^{S} D_{t+1}(b_u, \alpha + \rho Y_t + e_s) \right|.$$

Then generate Y_{t+1}^* by randomly sampling s^* from $\{1, \ldots, S\}$ and setting

$$Y_{t+1}^* = \alpha + \rho Y_t^* + e_{s^*}.$$

References

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

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

Ahn, Seung C and Peter Schmidt. 1995. "Efficient estimation of models for dynamic panel data." *Journal of Econometrics* 68 (1):5–27.

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." .

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.

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

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, Hamish Low, and Ian Preston. 2013. "Decomposing changes in income risk using consumption data." *Quantitative Economics* 4 (1):1–37.

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.

Dhaene, Geert and Koen Jochmans. 2015. "Split-panel jackknife estimation of fixed-effect models." *The Review of Economic Studies* 82 (3):991–1030.

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.

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* .

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.

Kaplan, Greg and Giovanni L Violante. 2010. "How much consumption insurance beyond self-insurance?" *American Economic Journal: Macroeconomics* 2 (4):53–87.

———. 2014. "A model of the consumption response to fiscal stimulus payments." *Econometrica* 82 (4):1199–1239.

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

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.

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.

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.

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.