Identification and estimation of dynamic random coefficient models

Wooyong Lee[*]

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

Abstract

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

[*] Economics Discipline Group, University of Technology Sydney. Address: 14-28 Ultimo Road, Ultimo, NSW 2007, Australia. Tel: (02) 9514 3074. Email: wooyong.lee@uts.edu.au

1 Introduction

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

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

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

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

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

Most research on random coefficient models with short panels focus on non-dynamic contexts (Chamberlain, 1992; Wooldridge, 2005; Arellano and Bonhomme, 2012; Graham and Powell, 2012), requiring that ε_{it} be uncorrelated with the entire history of regressors. This implies that future values of regressors are uncorrelated with current outcomes, which is difficult to justify. For example, a firm's labor purchase decision next year might correlate with this year's output since the firm might learn about its own efficiency of labor from the output. A researcher might also be interested in the dynamics itself. For example, earnings persistence of a household is an important parameter since high earnings persistence makes earnings shocks last, which reduces a household's consumption smoothing ability and hence household welfare.

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

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

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

Second, I show that the sharp bounds can be computed fast and reliably by exploiting the linear structure of the model. Computing sharp bounds obtained from dual representation involves solving a nested optimization problem in which a researcher maximizes an objective function that contains another minimization problem. An important computational issue is that the inner minimization problem is a global minimization problem of a possibly non-convex function, for which standard global optimization procedures are infeasible because the problem is nested and hence must be solved many times with precision. I show that for random coefficient models, the inner objective function is a polynomial. I then use a fast and exact algorithm to solve the global polynomial optimization problem, the semidefinite relaxation algorithm (Lasserre, 2010, 2015). Using this algorithm, sharp bounds for parameters of interest can be estimated timely, and inferences about the bounds based on testing moment inequalities can also be performed computationally tractably (Chernozhukov, Lee, and Rosen, 2013; Romano, Shaikh, and Wolf, 2014; Chernozhukov, Chetverikov, and Kato, 2019; Bai, Santos, and Shaikh, 2019). For researchers interested in using the semidefinite relaxation approach to global polynomial optimization, I offer a general-purpose R package optpoly that implements the approach[1].

Third, I estimate a reduced-form lifecycle model of earnings dynamics. Lifecycle earnings process serves as a key input in various economic models, including models of lifecycle consumption dynamics (Hall and Mishkin, 1982; Blundell, Pistaferri, and Preston, 2008; Blundell, Pistaferri, and Saporta-Eksten, 2016; Arellano, Blundell, and Bonhomme, 2017). Specifying an earnings process that highlights features of real data is important for calibrating and drawing conclusions from these models. This paper investigates unobserved heterogeneity in earnings of U.S. households in Panel Study of Income Dynamics (PSID) dataset. Guvenen (2007, 2009) pointed out that, when allowing for unobserved heterogeneity in time trend of earnings[2], earnings persistence is estimated to be significantly smaller than what is reported earlier. This paper estimates a more general model that also allows for unobserved heterogeneity in the earnings persistence itself. I find evidence of unobserved heterogeneity in earnings persistence, and the average persistence is estimated to be smaller than what is reported earlier.

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

The remainder of this paper is structured as follows. Section 2 introduces a dynamic random coefficient model. Section 3 and Section 4 present identification results about the model. Section 5 and Section 6 introduce estimation, inference and computation methods

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

[2] The resulting model is called heterogeneous income profiles (HIP) in the literature.

for parameters of interest. Section 7 checks performance of inference methods by simulation. Section 8 applies the methods to lifecycle earnings dynamics. Section 9 concludes. All tables are presented at the end of the paper.

2 Model and motivating examples

The dynamic random coefficient model is specified as:

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

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

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

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

The model is studied in a short panel context, which corresponds to the asymptotics that the number of individuals $N \to \infty$, but T is fixed. The random variables (γ_i, β_i), the random coefficients, have the same dimensions as (Z_{it}, X_{it}), and they can freely correlate among themselves and to (Z_i, X_{i1}). (γ_i, β_i) are viewed as unobserved random variables that are i.i.d. across i with a common nonparametric distribution, which is the sense that a random coefficient model extends a fixed effects model.

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

This paper considers parameter θ that has the form:

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

for some known function m. Theoretical results are presented for a generic function m, but I focus on the case in which m is either a polynomial or an indicator function with respect to V_i which allow computationally feasible estimation and inference procedures.

This choice of m includes many important parameters of interest. For example, θ can be an element of the mean of random coefficients $\mathbb{E}(\beta_i)$ or an element of the second moments $\mathbb{E}(\beta_i\beta_i')$. θ can also be the error variance $\mathbb{E}(\varepsilon_{it}^2)$ because

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

is a quadratic polynomial in (γ_i, β_i). Another example of θ is the CDF of β_i evaluated at b, in which m is the indicator function $\mathbf{1}(\beta_i \leq b)$ for some b:

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

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

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

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

The AR(1) model is a popular choice for empirical specification of the lifecycle earnings process, with Y_{it} being the log-earnings net of demographic variables, which is an important input in the lifecycle model of consumption and savings behavior[3]. Specification of the earnings process has a first-order influence on the model outcome. Persistence of earnings (β_i) governs earnings risk experienced by households, which is a fundamental motive of precautionary savings. The literature usually models it as an AR(1) process with no coefficient heterogeneity (Lillard and Weiss, 1979; Blundell, Low, and Preston, 2013; Gu and Koenker, 2017) or more simply as a unit root process, which is an AR(1) process with $\gamma_i = 0$ and $\beta_i = 1$ (Hall and Mishkin, 1982; Meghir and Pistaferri, 2004; Kaplan and Violante, 2014).

Guvenen (2007, 2009) estimated a variation of (3) where $\beta_i = \beta$ is homogeneous and a time trend is included. He pointed out that β is estimated to be significantly less than 1 when the time trend is allowed to be heterogeneous, in contrast to earlier findings that β is estimated to be close to 1. I find during application that, when β_i is allowed to be heterogeneous, $\mathbb{E}(\beta_i)$ is estimated to be significantly less than 1 regardless of whether the time trend is heterogeneous or not. Other studies that allow for coefficient heterogeneity in earnings include Browning, Ejrnaes, and Alvarez (2010) and Alan, Browning, and Ejrnæs (2018), with factor structure on the coefficients. □

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

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

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

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

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

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

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

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

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

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

which is a multivariate extension of (2).

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

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

time $t+1$ earnings equation, I obtain multivariate random coefficient model:

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

This can be written in matrix form:

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

□

3 Identification of means

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

Consider identifying a parameter that has the form:

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

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

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

3.1 Failure of point-identification

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

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

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

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

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

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

Proof. See Appendix A.1. □

The same result holds when $(Y_{i0}, Y_{i1}, Y_{i2}, \gamma_i, \beta_i)$ is discrete and the number of support points of (γ_i, β_i) is not too small relative to that of (Y_{i0}, Y_{i1}, Y_{i2}). Chamberlain (1993) showed that $\mathbb{E}(\beta_i)$ is not point-identified in (5) when Y_{it}s are discrete and ε_{it} is mean independent of Y_i^{t-1}. Proposition 1 generalizes the result, showing that point-identification also fails with stronger assumptions and continuous data. The proof suggests that the result holds for any finite $T \geq 2$. Failure of point-identification in both discrete and continuous cases in (5) suggests that it is a general feature of dynamic random coefficient models.

Proof of Proposition 1 has the following implication which is worth stated separately:

Corollary 1. *Under assumptions of Proposition 1, $\mathbb{E}(\beta_i)$ is point-identified if and only if there exists an unbiased estimator of β_i in individual time series, that is, a function $S^*(Y_{i0}, Y_{i1}, Y_{i2})$ which is a linear functional on the space of finite and countably additive signed Borel measures that are absolutely continuous with respect to the Lebesgue measure such that*

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

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

Proof. See Appendix A.2. □

I then show that there is no such S^*, which proves Proposition 1. The intuition for Corollary 1 is as follows. Since the distribution of β_i is unrestricted, information on individual β_i can only be obtained from its individual time series. In long panel context, such information can be obtained by a time series estimator of β_i that is consistent as $T \to \infty$. In short panel context, however, such estimator is not reliable because T is finite. Corollary 1 shows that a time series estimator that is unbiased for finite T is the only reliable information on β_i in short panels for point-identification.

3.2 Partial identification

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

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

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

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

I concisely write (1) and (2) defining $R_{it} = (Z'_{it}, X'_{it})'$ to be the vector of regressors at time t:

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

and

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

Recall that the parameter of interest, μ_e, is:

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

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

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

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

Consider the following assumptions:

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

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

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

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

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

Assumption 3 considers a specific choice of unconditional moment restrictions that are implications of (7). The first equation in Assumption 3 states that the "explained term" $(R_{it}'V_i)$ and the "error term" (ε_{it}) are orthogonal. The second equation states that ε_{it} is orthogonal to the full history of Z_{it} and the current history of X_{it}.

The following theorem shows that μ_e is partially identified under Assumptions 1 to 3 and additional regularity conditions. This theorem is a special case of Theorem 2 presented in the next section.

Theorem 1. *Suppose that Assumptions 1 to 3 hold, and assume additional regularity conditions which will be stated as Assumption 5 in the next section. In addition, assume that W_i are absolutely continuous with respect to the Lebesgue measure. Let $\lambda_t \in \mathbb{R}$ for $t = 1, \ldots, T$, and let μ_t be a real vector whose dimension is the same as $H_{it} = (Z_i{}', X_i^{t}{}')'$ for $t = 1, \ldots, T$. Also, let $\lambda \equiv (\lambda_1, \ldots, \lambda_T)$ and $\mu \equiv (\mu_1', \ldots, \mu_T')$. Then $L \leq \mu_e \leq U$ where*

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

and

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

where

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

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

Proof. See Appendix A.3. □

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

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

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

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

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* $\mathcal{R}_i^{-1}e$ *and* $\mathcal{R}_i^{-1}\mathcal{Y}_i$ *are degenerate across individuals, respectively.* $[\tilde{L}, \tilde{U}]$ *are the sharp bounds of* μ_e *under the following implication of Assumption 3:*

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}((R_{it}'V_i)\varepsilon_{it}) &= 0,\\
\sum_{t=1}^{T}\mathbb{E}(R_{it}\varepsilon_{it}) &= 0.
\end{aligned} \tag{9}$$

Proof. See Appendix A.4. □

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

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

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

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

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

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

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

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

and

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

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

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

Considering expectation on both sides yields

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

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

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

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

4 Identification of higher order moments and CDFs

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

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

for some known function $m : \mathcal{W} \times \mathcal{V} \mapsto \mathbb{R}$. I consider a generic set of unconditional moment restrictions:

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

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

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

ε_{it} does not appear in Assumption 4 because ε_{it} is understood as a deterministic function of (W_i, V_i) by the relationship $\varepsilon_{it} = Y_{it} - R_{it}' V_i$. Assumption 4 can also be considered as a generic moment equality model without connection to random coefficient models. A more general case that involves conditional moment restrictions is studied in Appendix B.

The following explains that the moment restrictions considered in the previous section is a special case of Assumption 4.

Example 4. Consider identification of $\mathbb{E}(e' V_i)$ discussed in the previous section. Assumption 3 implies the following moment functions. The ϕ_ks for $k = 1, \ldots, T$ are

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} \min_{P \in \mathcal{M}_{W\times V},\, P \geq 0} \int m(w,v)dP \qquad & \text{subject to} \\ & \int \phi_k(w,v)dP = 0, \quad k = 1, \dots, K, \\ & \int P(w, dv) = P_W(w) \text{ for all } w \in \mathcal{W}. \end{aligned} \tag{10}$$

The constraint $\int dP = 1$ is omitted since it is redundant given the last line of (10). $\int dP_W(w) = 1$ because it is a probability distribution.

(10) is a linear program (LP) in P, with the caveat that P is an infinite-dimensional object. (10) is not a tractable characterization of L in the sense that the estimation methods that it implies are computationally infeasible for random coefficient models. For example, (10) can be solved by discretizing the space of (W_i, V_i) and solving the discretized problem (Honoré and Tamer, 2006; Gunsilius, 2019), which is computationally infeasible for random coefficient models because the dimension of (W_i, V_i) is large. W_i contains the full history of regressors and response variables and V_i contains all random coefficients. For the random coefficient model with R regressors and T waves, P is a distribution on a $(RT + R + T)$-dimensional space.

My approach is to use the dual representation of (10) obtained by the duality theorem for infinite-dimensional LP (Galichon and Henry, 2009; Schennach, 2014). I assume the following regularity conditions:

Assumption 5. The following conditions hold.

(i) $\mathcal{W} \times \mathcal{V}$ is a compact set in an Euclidean space.

(ii) $(m, \phi_1, \ldots, \phi_K)$ are bounded Borel measurable functions on $\mathcal{W} \times \mathcal{V}$.

(iii) The set

$$D = \left\{ \left(\int \phi_1 dP, \ldots, \int \phi_K dP, \int P(\cdot, dv), \int m dP \right) \,\middle|\, P \in \mathcal{M}_{W \times V}, P \geq 0 \right\} \subseteq \mathbb{R}^K \times \mathcal{M}_W \times \mathbb{R}$$

is closed.

A sufficient condition for Assumption 5 (iii) is that the joint distribution of (W_i, V_i) in the data generating process, or its observationally equivalent one, is strictly positive on $\mathcal{W} \times \mathcal{V}$ (Anderson, 1983, Theorem 9).

The following theorem characterizes I using the dual representation of (10) and the corresponding problem for the sharp upper bound.

Theorem 2. *Suppose Assumptions 4 and 5 hold. Let $\lambda_k \in \mathbb{R}$ for $k = 1, \ldots, K$. Then $I = [L, U]$ where:*

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

and

$$U = \min_{\lambda_1, \ldots, \lambda_K} \mathbb{E} \left[\max_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} \right]. \tag{12}$$

Proof. See Appendix A.5. □

Proof of Theorem 2 has the following implication about point-identification of θ that is worth stated separately.

Corollary 2. *Suppose that assumptions of Theorem 2 hold. Suppose also that (W_i, V_i) are absolutely continuous with respect to the Lebesgue measure and that their joint density is strictly positive on $\mathcal{W} \times \mathcal{V}$. Then θ is point-identified if and only if there exists a function S^* which is a linear functional on $\mathcal{M}_W$ and real numbers $\lambda_1^*, \ldots, \lambda_K^* \in \mathbb{R}$ such that:*

$$m(W_i, V_i) + \sum_{k=1}^{K} \lambda_k^* \phi_k(W_i, V_i) = S^*(W_i)$$

almost surely on $\mathcal{W} \times \mathcal{V}$. When such S^* exists, θ is identified by $\theta = \mathbb{E}(S^*(W_i))$.

Proof. See Appendix A.6. □

Corollary 2 states that θ is point-identified if and only if the "Lagrangian" reduces to a function of data only. S^* can be thought of as an "unbiased" estimator because the term $\sum_{k=1}^{K} \lambda_k^* \phi_k(W_i, V_i)$ has zero expectation.

Theorem 2 and Corollary 2 do not explicitly involve dynamic random coefficient models. Theorem 2 is a general duality result for models of moment equalities, where the moment functions contain both observables and unobservables (Galichon and Henry, 2009; Schennach, 2014; Chesher and Rosen, 2017; Li, 2018). In general, it is not obvious that Theorem 2 leads to computationally feasible estimation and inference procedures. I show in the next sections that, for dynamic random coefficient models, I can obtain a computationally tractable estimation and inference procedure by exploiting that it is a linear model.

5 Estimation and inference

This section explains estimation and inference procedure for the identified sets discussed in the previous sections, focusing on describing the procedure. The next section discusses computation of the objects involved in the procedure.

5.1 Estimation

Theorem 2 characterizes the lower and upper bounds 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 is to replace expectations in (11) and (12) with sample means (the plug-in principle). Define $\hat{L}$ as an estimator for L:

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

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

$$\hat{U} = \min_{\lambda_1,\ldots,\lambda_K} \frac{1}{N}\sum_{i=1}^{N} \max_{v\in\mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} \equiv \min_{\lambda\in\mathbb{R}^K} \frac{1}{N}\sum_{i=1}^{N} G_U(\lambda, W_i), \tag{14}$$

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

The plug-in bound is used as a key object for estimation and inference, but the bound is not obvious to compute. For example, computation of (13) requires solving two types of optimization problems — solving the inner minimization problem over $\mathcal{V}$ and solving the outer maximization problem with respect to $\lambda_1, \ldots, \lambda_K$. Each problem has its own difficulties:

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

- The outer maximization problem must be solved globally, and it might be an optimization over a large dimensional space.

It will also be shown that, for computational tractability of the outer problem, the inner problem must be solved not only fast but also *exactly*. Thus, general-purpose global minimization methods are not computationally feasible except for low-dimensional cases, such as discrete space or $\mathbb{R}^2$.

The next section discusses how to deal with these computational issues. In this section, I discuss estimation and inference assuming that we can solve the two optimization problems numerically.

In what follows, I show consistency of lower plug-in bound in (13) to population lower bound in (11). The consistency of upper plug-in bound follows by the same argument.

In (13), the solution function of the inner optimization problem

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

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

$$\hat{L} = \max_{\lambda} \hat{L}(\lambda) = \max_{\lambda} \frac{1}{N} \sum_{i=1}^{N} G_L(\lambda, W_i) \tag{15}$$

as an estimator for

$$L = \max_{\lambda} L(\lambda) = \max_{\lambda} \mathbb{E}\left(G_L(\lambda, W_i)\right). \tag{16}$$

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

Proposition 3. *$G_L(\lambda, W_i)$ is globally concave in λ, which implies global concavity of $\hat{L}(\lambda)$.*

Proof. See Appendix A.7. □

Proposition 4. *Suppose that L exists and is finite, and that* $\operatorname{argmax}_{\lambda} L(\lambda)$ *is in the interior of $\mathbb{R}^K$. $\hat{L}$ then converges to L in probability.*

Proof. See Appendix A.8. □

5.2 Inference

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

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

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

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

which implies

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

These imply the following moment inequality conditions:

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

(17) is a moment inequalities model with infinite number of moment restrictions (indexed by $\lambda \in \mathbb{R}^K$). For computational tractability, I choose finite number of moment inequalities from (17). Let Λ_F be a finite subset of $\mathbb{R}^K$. Consider a moment inequalities model with parameter θ_F:

$$\begin{aligned} \mathbb{E}(G_L(\lambda, W_i) - \theta_F) \leq 0 \quad &\text{for all } \lambda \in \Lambda_F, \\ \mathbb{E}(\theta_F - G_U(\lambda, W_i)) \leq 0 \quad &\text{for all } \lambda \in \Lambda_F. \end{aligned} \tag{18}$$

Since Λ_F is a subset of $\mathbb{R}^K$, I can use (18) to make a conservative inference about θ in (17).

How conservative (18) is relative to (17) depends on how much information is contained in (18) relative to (17). Formal analysis of comparison between (17) and (18) is beyond the scope of this paper[5], but two observations provide guidance on how to choose an informative Λ_F in practice. First, the inequalities in (17) bind at two λ values, namely $\lambda_L^* = \text{argmax}_\lambda \, \mathbb{E}(G_L(\lambda, W_i))$ and $\lambda_U^* = \text{argmin}_\lambda \, \mathbb{E}(G_U(\lambda, W_i))$, and the inequalities are loose for λs that are distant from them (because G_L is concave and G_U is convex). This means that most of information in (17) is contained in the neighborhood of λ_L^* and λ_U^*[6]. Second, concavity of G_L (and convexity of G_U) implies that G_L and G_U are continuous, which means that considering a finite set of points in the neighborhood of λ_L^* and λ_U^* does not lead to serious loss of information compared to considering all points in the neighborhood. These observations lead to a practical strategy for choosing an informative Λ_F: estimate λ_L^* and λ_U^* using (13) and (14) and select a finite number of points in the neighborhoods of them. I check performance of this strategy by simulation in Section 7.

Given Λ_F, (18) is a standard moment inequalities model although Λ_F can be a large set. The literature on many moment inequalities (Romano, Shaikh, and Wolf, 2014; Bai, Santos, and Shaikh, 2019; Chernozhukov, Chetverikov, and Kato, 2019) propose procedures for computing a confidence interval $[L_\alpha, U_\alpha]$ that satisfies the following for large Λ_F:

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

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

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

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

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

where

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

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

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

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

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

$$T_{CCK}(\theta_F) \le c_{CCK}(\alpha). \tag{19}$$

In general, the procedure requires grid search over θ_F satisfying (19). In our case, two observations facilitate efficient search of θ_F. First, $c_{CCK}(\alpha)$ does not depend on θ_F, because c_{CCK} does not. This means $c_{CCK}(\alpha)$ is computed only once and is fixed when evaluating (19) for different θ_F. Second, if $\sigma_{G_L}(\lambda), \sigma_{G_U}(\lambda) > 0$ for all $\lambda \in \Lambda_F$ and the confidence interval is not empty, $T_{CCK}(\theta_F)$ is convex in θ_F. These observations lead to the following efficient algorithm for computing a confidence interval. First, perform a crude grid search to find any interior point of the confidence interval, i.e. a point θ_F^* such that $T_{CCK}(\theta_F^*) < c_{CCK}(\alpha)$. Then solve the equation $T_{CCK}(\theta) = c_{CCK}(\alpha)$ on $(-\infty, \theta_F^*)$ to find L_α, and on (θ_F^*, ∞) to find U_α.

The inference procedure naturally extends to a vector-valued parameter $\theta \in \mathbb{R}^d$, by considering (17) for every entry of θ. For example, the moment inequalities for $\theta =$

$(\theta_1, \theta_2) \in \mathbb{R}^2$ is:

$$\begin{aligned} \mathbb{E}(G_{L1}(\lambda, W_i) - \theta_1) \leq 0 \quad & \text{for all } \lambda \in \mathbb{R}^K, \\ \mathbb{E}(\theta_1 - G_{U1}(\lambda, W_i)) \leq 0 \quad & \text{for all } \lambda \in \mathbb{R}^K, \\ \mathbb{E}(G_{L2}(\lambda, W_i) - \theta_2) \leq 0 \quad & \text{for all } \lambda \in \mathbb{R}^K, \\ \mathbb{E}(\theta_2 - G_{U2}(\lambda, W_i)) \leq 0 \quad & \text{for all } \lambda \in \mathbb{R}^K, \end{aligned} \tag{20}$$

where G_{Uk} and G_{Lk} are G_L and G_U in (17) for θ_k, $k = 1, 2$. Inference can then be performed by the same procedure, giving a confidence region in $\mathbb{R}^2$. This extension can be used to compute a confidence interval for the variance of random coefficients which involves the first and the second moments.

5.3 Estimation under empty plug-in bound

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

There are two approaches for dealing with this issue. First, a researcher may obtain a point estimate that minimizes distance between the model and the data. Second, a researcher may directly obtain a confidence interval, not insisting a point estimate. I propose using the second approach for dynamic random coefficient models. For completeness, I discuss the first approach in this subsection, and the next subsection discusses the second approach.

In GMM estimation, a researcher defines an estimator as the parameter value that minimizes the GMM criterion that may be strictly positive. A similar approach can be used for dynamic random coefficient models in two steps[7]. In the first step, the researcher finds the smallest $\delta \geq 0$ that satisfies:

$$|\mathbb{E}(\phi_k(W_i, V_i))| \leq \delta, \quad k = 1, \ldots, K. \tag{21}$$

This can be thought of as an absolute-value GMM objective. The following proposition

[7] Andrews and Kwon (2019) study and formalize this approach for standard GMM estimation of moment equality models without unobservables.

explains how to compute the smallest δ.

Proposition 5. *Given the sample $(W_1, \dots, W_N)$, consider linear programming problem:*

$$\min_{P \in \mathcal{M}_{W \times V},\, P \geq 0,\, \delta \geq 0} \delta \qquad \textit{subject to} \qquad \begin{aligned} &\left| \int \phi_k(W_i, V_i) dP \right| \leq \delta, \quad k = 1, \dots, K, \\ &\int P(w, dV_i) = \hat{P}_W(w) \ \textit{for all } w \in \mathcal{W}, \end{aligned} \tag{22}$$

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

$$\max_{\lambda_1, \dots, \lambda_K} \frac{1}{N} \sum_{i=1}^{N} \min_{v \in \mathcal{V}} \left\{ \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} \qquad \textit{subject to} \qquad \sum_{k=1}^{K} |\lambda_k| \leq 1. \tag{23}$$

Proof. See Appendix A.9. □

Proposition 5 shows that a researcher can find the smallest δ by solving (23), which is similar to the plug-in bound problem. A difference is that (23) is a constrained optimization problem, but the constraint has a simple structure whose Jacobian can be derived in closed-form.

Let δ^* be the solution to (23), and let $\delta \geq \delta^*$. The second step then computes modified plug-in bounds. I compute the lower bound with negative L^1 penalty on λ, with δ being the penalty multiplier:

$$\hat{L}_{pen} = \max_{\lambda_1, \dots, \lambda_K} \left[\frac{1}{N} \sum_{i=1}^{N} \min_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} - \delta \sum_{k=1}^{K} |\lambda_k| \right]. \tag{24}$$

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

Proposition 6. *Given the sample $(W_1, \dots, W_N)$ and given $\delta \in \mathbb{R}$, consider the linear programming problem:*

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

where $\hat{P}_W$ is the empirical distribution of W_i constructed from $(W_1, \dots, W_N)$. Its solution then equals $\hat{L}_{pen}$, defined in (24).

Proof. See Appendix A.10. □

Proposition 6 shows that (24) equals the smallest value of θ among the distributions whose absolute-value GMM objective defined in (21) is at most δ. In principle, such a distribution is not necessarily unique even when $\delta = \delta^*$. If it is unique, the modified plug-in bound becomes a point.

In practice, due to machine precision or the stopping criterion of optimization methods, the numerical solution to (23) might be strictly smaller than the analytical solution δ^*. To resolve this problem, a researcher may choose δ to be larger than δ^*, in which case (24) computes the smallest value of θ among the distributions that attain the *near-minimum* of the absolute-value GMM criterion. In the case that the minimizer distribution is unique, the modified plug-in bounds with $\delta > \delta^*$ becomes a small interval instead of a point.

Although (24) resolves empty plug-in bound problem, it has two drawbacks. First, it is an ad-hoc approach, having no formal justification for why relaxation of moment conditions is a good idea. Second, the procedure may yield a point (or a small interval) as the identified set for the parameter of interest even if the model is partially identified. The literature dealt with the second problem by choosing δ that is reasonably larger than δ^* (Mogstad, Santos, and Torgovitsky, 2018), but how much larger it should be remains a question. The next subsection discusses a more principled approach, which is directly computing a confidence interval without insisting a point estimate.

5.4 Inference under empty plug-in bound

This subsection proposes an inference procedure that is valid even if the plug-in bound is empty. The idea can be understood again by comparing it to GMM estimation, in which a researcher can use the point estimate to construct a valid confidence interval even if the point estimate attains a strictly positive GMM objective. I take the same approach, using (24) as a building block for the inference procedure.

Note that the proposed inference procedure that tests (18) does not involve the plug-in bound per se. The plug-in bound is involved only in the step of choosing Λ_F, which I propose to be the set of λs that are close to the solutions to plug-in bound problems. The inference procedure is valid regardless of whether the plug-in bound is empty or not; the issue is that there is no guidance for choosing Λ_F when the plug-in bound is empty. In what follows, I propose a strategy for choosing Λ_F when the plug-in bound is empty.

I propose using the previous subsection as a building block for choosing Λ_F. The procedure consists of three steps. The first step solves (23) and finds minimum δ^*. The second step considers a grid of positive real numbers $\{\delta_1, \ldots, \delta_M\}$ such that $\delta_m \geq \delta^*$ for all $m \in \{1, \ldots, M\}$. Then, for each δ_m, compute the penalized plug-in bounds:

$$\begin{aligned} \tilde{\lambda}_L(\delta_m) &= \underset{\lambda_1,\ldots,\lambda_K}{\operatorname{argmax}} \left[\frac{1}{N} \sum_{i=1}^{N} \min_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} - \delta_m \sum_{k=1}^{K} |\lambda_k| \right]. \\ \tilde{\lambda}_U(\delta m) &= \underset{\lambda_1,\ldots,\lambda_K}{\operatorname{argmin}} \left[\frac{1}{N} \sum_{i=1}^{N} \max_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} + \delta_m \sum_{k=1}^{K} |\lambda_k| \right]. \end{aligned} \tag{26}$$

The third step then chooses points in the neighborhoods of *every* $\tilde{\lambda}_L(\delta_m)$ and $\tilde{\lambda}_U(\delta_m)$. During simulation and application, I choose points by adding Gaussian noise to every $\tilde{\lambda}_L(\delta_m)$ and $\tilde{\lambda}_U(\delta_m)$ whose standard deviations are inversely proportional to the gradients of the $\tilde{\lambda}$s. I check performance of this approach by simulation in Section 7.

When $\delta^* = 0$, i.e. when the plug-in bound is non-empty, a researcher may choose $M = 1$ with $\delta_1 = 0$, in which case the inference procedure reduces to the procedure in Section 5.2 under non-empty plug-in bound. This means that the inference procedure discussed in this subsection generalizes the procedure discussed in Section 5.2.

6 Computation

This section discusses computation of objects involved in estimation and inference. In particular, the discussion focuses on computation of the two optimization problems in the plug-in lower bound in (13), which apply similarly to all the other objects such as plug-in upper bound in (14), moment inequalities in (18) and penalized plug-in bounds in (26).

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

6.1 The inner problem

The inner optimization problem of (13) is to evaluate the function

$$G_L(\lambda, w) = \min_{v \in \mathcal{V}} \left\{ m(w, v) + \sum_{k=1}^{K} \lambda_k \phi_k(w, v) \right\} \tag{27}$$

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

One difficulty in evaluating G_L is that the minimization problem must be solved globally. In the simple case that $\mathcal{V}$ is discrete or low-dimensional, 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, these are not likely to be the case.

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

Example 5. In Section 3, I showed identification of the mean parameter $\mathbb{E}(e'V_i)$ under Assumptions 1 to 3. In this case, the m function is given by

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

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

$$(R_{it}'V_i)(Y_{it} - R_{it}'V_i), \quad t = 1, \dots, T, \tag{28}$$

and the entries of the vectors

$$(Z_i', X_i^{t\prime})'(Y_i - R_{it}'V_i), \quad t = 1, \dots, T, \tag{29}$$

which are at most second-order polynomials of V_i. These moment functions are what I use during application for computing confidence intervals of means of random coefficients.

Example 6. Consider identification of an element of $\mathbb{E}(V_iV_i')$. Then m is an element of V_iV_i', which is a second-order polynomial of V_i. Consider the moment restriction $\mathbb{E}((R_i'V_i)^3\varepsilon_{it}) = 0$, in which case the ϕ_ks consist of the functions

$$(R_{it}'V_i)^3(Y_{it} - R_{it}'V_i), \quad t = 1, \dots, T, \tag{30}$$

which are fourth-order polynomials of V_i. A researcher may also consider the moment functions in Assumption 3, in which case he/she sets the additional ϕ_ks to be (28) and (29). These moment functions are what I use during application for computing confidence intervals of variances 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 obtaining finite bounds, I choose ϕ_ks so that the inner objective function is an even order polynomial whose order is strictly larger than the order of the parameter of interest. In Examples 5 and 6, I choose (28) to obtain a second order polynomial and (30) to obtain a fourth order polynomial as inner objectives. The inner objective function then has its leading coefficient positive or negative depending on the signs of λ, which leads to finite inner miminum or maximum that yields finite lower and upper bounds.

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

This extension is useful 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 $v^0 \in \mathbb{R}$. Consider identification of the CDF of V_{i1} evaluated at v^0. I set m to be

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

which is an indicator function of V_i. Consider the same set of moment restrictions as Example 5, in which case the ϕ_ks are at most second-order polynomials in V_i. The m function then partitions $\mathcal{V}$ into two exclusive sets $\mathcal{V}_1 = \{(v_1, \ldots, v_{q+p}) \mid v_1 \leq v\}$ and $\mathcal{V}_2 = \{(v_1, \ldots, v_{q+p}) \mid v_1 > v\}$, and $m = 1$ on $\mathcal{V}_1$ and $m = 0$ on $\mathcal{V}_2$. The inner objective function is then a second-order polynomial within each of $\mathcal{V}_1$ and $\mathcal{V}_2$, for which I can compute the minimum. I can then evaluate the inner objective by taking the smaller optimum between those in $\mathcal{V}_1$ and $\mathcal{V}_2$.

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

6.1.1 Global optimization of quadratic polynomials

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

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

where A is a $\dim(v) \times \dim(v)$ symmetric matrix, b is a $\dim(v)$-dimensional vector, and $c \in \mathbb{R}$. If the inner objective of (27) is expressed in this standard form, (A, b, c) are functions of w, the data.

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

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

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

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

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

$$\min_{v \in \mathcal{V}} Q(v) = c - \frac{1}{4}b'A^{-1}b. \tag{31}$$

If A is not positive definite, the minimum of Q is negative infinity unless A has a zero eigenvalue. If A has a zero eigenvalue, Q has a finite minimum if the first-order condition

$$2Av + b = 0$$

has an infinite number of solutions for which the value of Q is the same. Otherwise, the minimum of Q is negative infinity.

In the context of (27), if the data w is from a continuous distribution, A has a zero eigenvalue with probability zero. Therefore, for continuous data, a researcher may rule out the possibility of zero eigenvalue in practice. That is, a researcher may simply use (31) to express (27) in a closed-form if and only if A is positive definite; otherwise the solution is negative infinity.

The heuristic method discussed in the above applies when $\mathcal{V}$ in (27) is unbounded. In some cases, a researcher may consider restricting $\mathcal{V}$ to be a bounded set, such as restricting the autoregressive parameter of the AR(1) model to be between 0 and 1. In that case, the heuristic method can be modified to incorporate the constraint using the Lagrange multiplier method. Alternatively, he/she may use quadratic optimization softwares.

6.1.2 Global optimization of generic polynomials

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

Consider computing 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 $u(v)u(v)'$. Let J be the cardinality of $\{p_j(v)\}$.

I can then express a fourth-order polynomial in standard form:

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

where a_j is the coefficient on the monomial $p_j(v)$.

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

$$\min_{P_V \in \mathcal{M}_V,\ \int dP_V = 1} \int \pi(v) dP_V \tag{32}$$

where P_V is a probability distribution on $\mathcal{V}$. (32) is minimized at the point-mass distribution concentrated at the minimizer of $\pi(v)$.

Since $\pi(v)$ is a linear combination of p_js, I can rewrite (32) 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:

$$\min_{M_1, \ldots, M_J \in \mathbb{R},\ M_1 = 1} \sum_{j=1}^{J} a_j M_j \qquad \text{subject to} \qquad M_j = \int p_j(v) dP_V \text{ for some } P_V \in \mathcal{M}_V. \tag{33}$$

Except for the fact that the constraint is complicated, (33) is a minimization over $\mathbb{R}^J$ and the objective is linear (and hence convex) in the choice variables.

The idea is then to replace the constraint in (33) with a convex constraint that only involves $(M_1, \ldots, M_J)$. The constraint in (33) tells that $(M_1, \ldots, M_J)$ must be moments of some underlying distribution. Checking this constraint relates to a problem called *the moment problem* in mathematics; "Given the sequence of real numbers $(M_1, \ldots, M_J)$, can they be justified as moments of some distribution?"

A sequence of real numbers must satisfy some relationship between them for them to be justified as moments. For example, a random variable X must satisfy:

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

because $\mathrm{Var}(X)$ has to be nonnegative. This is equivalent to condition:

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

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

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

If $(M_1, \ldots, M_J)$ are moments, then it must be:

$$\mathcal{L}(u(v)u(v)') \text{ is positive semidefinite} \tag{34}$$

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

(34) is a convex constraint because the set of positive semidefinite matrices is a convex set in the space of vectorized matrix entries. Therefore, if I replace the constraint in (33) with (34), I obtain a convex optimization problem:

$$\min_{M_1, \ldots, M_J \in \mathbb{R}} \sum_{j=1}^{J} a_j M_j \qquad \text{subject to} \qquad \mathcal{L}(u(v)u(v)') \text{ is positive semidefinite.} \tag{35}$$

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

The SDP approach to polynomial optimization solves (35), the *semidefinite relaxation*, which can be solved fast and reliably using SDP solvers available in the industry. The algorithm offers *certificate of optimality*, a condition for the optimal value of $(M_1, \ldots, M_J)$, satisfying which means that the solution to (35) equals the global optimum. For researchers interested in using the semidefinite relaxation approach to global polynomial optimization, I offer a general-purpose R package optpoly that implements the approach[8]. Alternatively, a Matlab package Gloptipoly (Henrion, Lasserre, and Löfberg, 2008) is also available.

The solution to (35) (i.e., the SDP solution) is less than or equal to the solution to (33) since a necessary condition is weaker than the original condition. The semidefinite relaxation approach solves a sequence of the SDP programs, or a *hierarchy* of the SDP programs, until the certificate of optimality is obtained, which is known to be obtained in a finite number of steps under suitable conditions. Even if a researcher does not solve the hierarchy of the SDPs, he/she can take an SDP solution as a lower bound for (33), in which case the resulting plug-in bound is a non-sharp and yet a valid bound for θ.

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

6.2 The outer problem

I turn to the outer optimization problem of (13). A researcher needs to solve the optimization problem:

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

Assume that the researcher can evaluate G_L exactly using the algorithms in the previous subsection. The remaining difficulty then is how to solve the optimization problem where K can be potentially large.

Recall that G_L is globally concave, as shown in Proposition 3. This implies that there is only one local maximum in the outer optimization problem, which is also the global maximum. Under suitable conditions, G_L is differentiable when $K = 1$ (Milgrom and Segal, 2002, Theorem 3), which can be extended to show that G_L is directionally differentiable for $K > 1$. This suggests that the researcher can solve the outer problem using fast convex

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

optimization algorithms such as gradient descent methods.

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

7 Simulation

This section examines by simulation the performance of the inference procedure discussed in Section 5. The simulation considers the AR(1) model given in (3) as the data generating process (DGP):

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

In DGP, $\gamma_i \in \mathbb{R}$ and $\beta_i \in [0, 1]$ follow Normal and Beta distributions respectively, and their joint distribution is given by Gaussian copula. ε_{it} follows an independent Normal distribution with mean zero and variance varying over t. Conditional on (γ_i, β_i) where $\beta_i \leq 0.9$, Y_{i0} is generated from the stationary distribution implied by (γ_i, β_i) and the variance of ε_{i1}:

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

In contrast, conditional on (γ_i, β_i) where $\beta_i > 0.9$, Y_{i0} is generated from an independent Normal distribution because the stationary distribution implied by (γ_i, β_i) produces extreme values when β_i is close to 1. Parameter values of the distribution of (γ_i, β_i) and ε_{it} are chosen based on the estimates of the income process in the application.

Simulation data are generated in two steps. The first step simulates a dataset of 100,000 observations from the parametric model described above. The second step then creates Monte Carlo samples from the 100,000 observations in the first step, by sampling observations with replacement. I consider the 100,000 observations as a *finite population* from which Monte Carlo samples are generated. Using a finite population is convenient because I can exactly compute the population identified set using (13) and (14), whereas computing the identified set of the parametric model itself is computationally infeasible.

Table 1 lists population bounds for $\mathbb{E}(\beta_i)$ computed from the finite population of size 100,000 with $T = 5, 10$ or 15. For each T, I compute the bounds using moment conditions:

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

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

For each (T, L), I create Monte Carlo replications by sampling $N = 750$ or 1000 observations from the finite population with replacement. I then compute confidence interval for $\mathbb{E}(\beta_i)$ in each Monte Carlo replication, using generalized inference procedure discussed in Section 5.4. The grid of $\{\delta_m\}$ is set to be

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

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

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

8 Application to lifecycle earnings dynamics

8.1 Overview

Lifecycle earnings dynamics is a key input in various macroeconomic models. For example, in models of consumption and savings dynamics (Hall and Mishkin, 1982; Blundell, Pistaferri, and Preston, 2008; Blundell, Pistaferri, and Saporta-Eksten, 2016; Arellano, Blundell, and Bonhomme, 2017), households facing higher risk in earnings dynamics accumulate more precautionary savings to smooth consumption. Households save a lot when they experience a positive earnings shock, which is used to maintain consumption during a negative earnings shock. Specifying earnings process that highlights features of real data is important for calibrating and drawing conclusions from these models.

When used as an input, it is common to specify earnings dynamics using a parsimonious linear model. It consists of permanent and transitory income processes[9]:

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

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

Guvenen (2007, 2009) studied two leading views on unobserved heterogeneity in earnings dynamics. Consider two earnings processes:

$$\begin{aligned} Y_{it} &= \alpha_i + z_{it} + \varepsilon_{it}, & z_{it} &= \rho z_{i,t-1} + \eta_{it}, \\ Y_{it} &= \alpha_i + \beta_i h_{it} + z_{it} + \varepsilon_{it}, & z_{it} &= \rho z_{i,t-1} + \eta_{it}, \end{aligned} \tag{36}$$

where h_{it} is potential years of experience and (α_i, β_i) are heterogeneous deviations from common trends. These two models are called Restricted Income Profiles (RIP) process and Heterogeneous Income Profiles (HIP) process, respectively. In both models, ρ a key parameter because it represents earnings risk that households face. The literature reports $0.5 < \rho < 0.7$ and $\text{Var}(\beta_i) > 0$ for HIP process (e.g. Lillard and Weiss, 1979; Baker, 1997), which means households experience modest persistence and heterogeneous trends. In contrast, MaCurdy (1982) tested hypothesis that $\text{Var}(\beta_i) = 0$ and did not reject the hypothesis. The literature reports $\rho \approx 1$ for RIP process (e.g. Abowd and Card, 1989; Topel and Ward, 1992), meaning households experience extreme persistence and homogeneous trends. Guvenen (2007) studied implications of the two models on consumption data and found that HIP is more consistent with features of consumption data. Guvenen (2009) pointed out that misspecifying HIP process as a RIP process leads to an upward biased estimator of ρ, obtaining $\rho \approx 1$.

While there is vast literature on heterogeneity in β_i and its influence on ρ, there is relatively little work on investigating heterogeneity in ρ itself. Recent studies on it include Browning, Ejrnaes, and Alvarez (2010) and Alan, Browning, and Ejrnæs (2018) who estimated a parametric model of income process in which heterogeneity in ρ is given by

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

a factor structure. In this section, I investigate heterogeneity in ρ by estimating generalization of (36) where $\rho = \rho_i$. I treat the modified model as a random coefficient model, meaning that distribution of ρ_i and its dependence to (α_i, β_i) and initial earnings Y_{i0} are unrestricted. Distributions of η_{it}s are also not restricted and may depend on ρ_i, allowing for heteroskedasticity.

In the remainder of this section, I find that, when $\rho = \rho_i$ is allowed to be heterogeneous, both RIP and HIP have similar estimates of $\mathbb{E}(\rho_i)$ that is significantly less than 1. Confidence intervals for $\mathbb{E}(\rho_i)$ in the two processes have substantial overlap, both having upper confidence limits around 0.6 at 90% confidence level. Confidence intervals for the CDF of ρ_i are also similar between the two processes. These results suggest that choosing RIP over HIP or vice versa may not lead to serious misspecification when ρ is allowed to be heterogeneous. I also find evidence of heterogeneity in ρ_i. For example, confidence interval for $\text{Var}(\rho_i)$ in RIP process has a lower confidence limit of 0.067 at 90% confidence level, implying a lower confidence limit of 0.258 for standard deviation of ρ_i.

8.2 Data

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

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

Note that potential experience is a strictly exogenous variable since it is a deterministic function of age.

The data used in Guvenen (2009) is an unbalanced panel. To create a balanced panel, I collect individuals with consecutive waves of data from 1976 to 1991, which yields $N = 800$ and $T = 15$ taking the first wave as an initial value of earnings.

Recall that my method requires that there is no multicollinearity in each individual time series (Assumption 2). To maintain this assumption, I removed 40 individuals (that is, five percent of data) who have smallest variations in their reported incomes, giving a dataset of $N = 760$ and $T = 15$. Estimation results with this removal does not qualitatively differ from the results with full data, which can be found in Appendix C.1.

8.3 From income processes to random coefficient models

To apply my method to income processes, I transform two models in (36) to random coefficient models. I first separate ε_{it} from Y_{it} using a simulation based de-noising method inspired by Arellano and Bonhomme (2018), described in Appendix C.2, obtaining simulated observations of $\tilde{Y}_{it}$:

$$\begin{aligned} \tilde{Y}_{it} &= \alpha_i + z_{it}, & z_{it} &= \rho z_{i,t-1} + \eta_{it}, \\ \tilde{Y}_{it} &= \alpha_i + \beta_i h_{it} + z_{it}, & z_{it} &= \rho z_{i,t-1} + \eta_{it}. \end{aligned} \tag{37}$$

Estimation results with de-noising does not qualitatively differ from the results without de-noising, which can also be found in Appendix C.2. I then use quasi-differencing to (37) and transform them to random coefficient models. Quasi-differencing the first model (RIP) yields:

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

Quasi-differencing the second model (HIP) yields:

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

These are standard random coefficient models. Note that h_{it} is strictly exogenous since it is a deterministic function of age.

8.4 Strategy for estimation and inference

For each model, I compute confidence intervals for $\mathbb{E}(\rho_i)$, $\mathrm{Var}(\rho_i)$ and $\mathbb{P}(\rho_i \leq r)$ for a grid of $r = 0, 0.1, \ldots, 0.9, 1$. For $\mathbb{E}(\rho_i)$ and $\mathbb{P}(\rho_i \leq r)$, I use moment restrictions stated in Example 5. In particular, I use, for the first model:

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

and for the second model:

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

These make the inner objective as a second order polynomial. I then solve the inner optimization problem as a closed-form.

I use additional moment restrictions to compute confidence interval for $\text{Var}(\rho_i)$. Additional moment conditions for the first model are:

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

and for the second model are:

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

The first additional moment restriction in both models was stated in Example 6, which makes the inner objective as a fourth order polynomial. The rest of additional moment restrictions then add low order terms to the inner objective. These additional restrictions yield finite lower and upper bounds on the second moments of $(\tilde{\alpha}_i, \tilde{\beta}_i, \rho_i)$. The inner problem is then solved using the SDP method with hierarchy of length two.

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

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

Then $P = 50$ points in the neighborhood of each $\tilde{\lambda}_L(\delta_m)$ and $\tilde{\lambda}_U(\delta_m))$ are sampled using additive Gaussian noise whose standard deviation is inversely proportional to the gradient at the penalized plug-in bounds, and the critical value is computed with 2000 multiplier bootstrap replications.

8.5 Results

Confidence intervals for $\mathbb{E}(\rho_i)$ and $\text{Var}(\rho_i)$ are given in Table 4. Both models estimate $\mathbb{E}(\rho_i)$ to be significantly less than 1. Moreover, their confidence intervals show substantial overlap, having similar upper confidence limits. These suggest that specifying homogeneous or heterogeneous β does not lead to serious misspecification when ρ is allowed to be heterogeneous. Confidence interval for $\text{Var}(\rho_i)$ suggest heterogeneity in ρ_i, having a lower confidence limit of 0.067 for RIP process implying standard deviation of 0.258.

Confidence intervals for the CDF of ρ_i for a grid of values are given in Table 5. They suggest substantial heterogeneity in ρ_i. For example, confidence intervals for RIP process suggest that at least 15% of households are estimated to have $\rho_i \leq 0.5$, while another at least 15% of households are estimated to have $\rho_i > 0.5$. Confidence intervals for the two CDFs show substantial overlap as well.

9 Conclusion

This paper studies identification and estimation of dynamic random coefficient models. I show that the model is not point-identified and yet partially identified, and I characterize a sharp identified set using the duality representation of infinite-dimensional linear programming. A computationally feasible estimation and inference procedure for the identified set that uses literature on testing moment inequalities is proposed. Computational feasibility of the procedure is achieved using a fast and exact algorithm for global polynomial optimization.

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

10 Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

———. 2018. "Recovering latent variables by matching." *Working paper* .

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

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

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

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

Blundell, Richard, 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.

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

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

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

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

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

Gu, Jiaying and Roger Koenker. 2017. "Unobserved heterogeneity in income dynamics: An empirical Bayes perspective." *Journal of Business & Economic Statistics* 35 (1):1–16.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Meghir, Costas and Luigi Pistaferri. 2004. "Income variance dynamics and heterogeneity." *Econometrica* 72 (1):1–32.

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.

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

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

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

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

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

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

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

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

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

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

Appendices

A Proofs

A.1 Proof of Proposition 1

In this proof, I take as granted the results of Appendix B, which generalizes the results of Section 4 of the main text. Note that the last condition of Assumption 7 in Appendix B is satisfied because the joint density of $(Y_{i0}, Y_{i1}, Y_{i2}, \gamma_i, \beta_i)$ is strictly positive (Anderson, 1983, Theorem 9).

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

Suppose that $\mathbb{E}(\beta_i)$ is point-identified, from which I draw contradiction. 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 linear functionals on the spaces of finite and countably additive signed Borel measures that are absolutely continuous with respect to the Lebesgue measure. Corollary 3 in Appendix B then implies that (f^*, g_1^*, g_2^*) must satisfy

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

almost surely on $\mathcal{C}_0^5$. Substituting $\varepsilon_{it} = Y_{it} - \gamma_i - \beta_i Y_{i,t-1}$ in (38) yields:

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

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

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

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

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

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

If not, there exists a subset of $\mathcal{C}_0^6$ with positive Lebesgue measure in which $\triangle_{\tilde{\gamma},\gamma} g_2^* \neq 0$, which implies that (40) is violated with positive measure (one can change the value of y_2

without changing $(\gamma, \tilde{\gamma}, \beta, y_0, y_1)$ within this subset).

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

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

Next, consider any $\gamma, \beta, \tilde{\beta}, y_0, y_1, y_2 \in \mathcal{C}$ such that $\beta \neq \tilde{\beta}$. I evaluate (39) at $(\gamma, \beta, y_0, y_1, y_2)$ and $(\gamma, \tilde{\beta}, y_0, y_1, y_2)$, and I take difference:

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

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. In (42), y_2 only appears in the third term. This implies that, almost surely:

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

(40) then simplifies to

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

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

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

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

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

(44) then simplifies to

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

which implies, since $\hat{\gamma} - \tilde{\gamma} = \tilde{\gamma} - \gamma \neq 0$:

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

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

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

I then apply the same argument to (42) with respect to β, which yields:

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

(39) then 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), so their coefficients on both sides must coincide:

$$g_1^* + g_2^* = 0,$$
$$-y_0 g_1^* - y_1 g_2^* = 1.$$

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

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

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

A.2 Proof of Corollary 1

This is a direct consequence of Corollary 3 in Appendix B. If $\mathbb{E}(\beta_i)$ is point-identified, Corollary 3 implies there exists (f^*, g_1^*, g_2^*) such that

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

almost surely on $\mathcal{C}_0^5$. Then it follows that $S^*(Y_{i0}, Y_{i1}, Y_{i2}) = f^*(Y_{i0}, Y_{i1}, Y_{i2})$ because:

$$\mathbb{E}(f^*(Y_{i0}, Y_{i1}, Y_{i2})|\beta_i) = \mathbb{E}\left(\beta_i - g_1^*(\gamma_i, \beta_i, Y_{i0})\varepsilon_{i1} - g_2^*(\gamma_i, \beta_i, Y_{i0}, Y_{i1})\varepsilon_{i2}|\beta_i\right) = \beta_i$$

where the second equality follows by $\mathbb{E}(\varepsilon_{i1}|\gamma_i, \beta_i, Y_{i0}) = 0$ and $\mathbb{E}(\varepsilon_{i2}|\gamma_i, \beta_i, Y_{i0}, Y_{i1}) = 0$.

Conversely, if there exists $S^*(Y_{i0}, Y_{i1}, Y_{i2})$ such that $\mathbb{E}(S^*(Y_{i0}, Y_{i1}, Y_{i2})|\beta_i) = \beta_i$, then:

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

which completes the proof. □

A.3 Proof of Theorem 1

This is a special case of Theorem 2 in Section 4 where the inner optimization problem is solved in closed-form for $\lambda < 0$ in the lower bound and for $\lambda > 0$ in the upper bound, using the method discussed in Section 6.1.1. Note that it suffices to consider $\lambda < 0$ in the lower bound because (i) the outer optimization problem is convex (Proposition 3), so it is not loss of generality to consider $\lambda < 0$ if the optimum is attained with $\lambda < 0$, and (ii) the optimum is attained with $\lambda < 0$ since the objective of the inner optimization problem, which is a quadratic polynomial, diverges to negative infinity when $\lambda \geq 0$ since its second order derivative is not positive definite, as discussed in Section 6.1.1. The same argument applies to considering $\lambda > 0$ in the upper bound. □

A.4 Proof of Proposition 2

For the proof, it suffices to show the sharpness of $[\tilde{L}, \tilde{U}]$ under (9). The inclusion $[L, U] \subseteq [\tilde{L}, \tilde{U}]$ then follows from inclusion of the assumptions.

In what follows, I show that $\tilde{U}$ is the sharp upper bound under (9). The same argument applies to $\tilde{L}$. This proof relies on Theorem 2 in Section 4.

By Theorem 2, the sharp upper bound is:

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

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

With the notation, I can write the sharp upper bound concisely as

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

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

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

I expand the terms and obtain:

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

I solve this problem with respect to μ for a fixed λ. The first order condition with respect to μ given λ 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:

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

I substitute this into (45) and obtain:

$$\begin{aligned}\min_{\lambda} \Big\{&[e-\lambda\mathbb{E}(\mathcal{Y}_i)]'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i) + \frac{1}{4\lambda}e'\mathbb{E}(\mathcal{R}_i^{-1})e + \frac{\lambda}{4}\mathbb{E}(\mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i) \\ &+ \frac{1}{4}\left[\frac{1}{\lambda}e'\mathbb{E}(\mathcal{R}_i)^{-1}e - 2e'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i) + \lambda\mathbb{E}(\mathcal{Y}_i)'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i)\right] \\ &+ \frac{1}{2}e'\mathbb{E}(\mathcal{R}_i^{-1}\mathcal{Y}_i) - \frac{1}{2}e'\mathbb{E}(\mathcal{R}_i)^{-1}\left[\frac{1}{\lambda}e - \mathbb{E}(\mathcal{Y}_i)\right] - \frac{1}{2}[e-\lambda\mathbb{E}(\mathcal{Y}_i)]'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i)\Big\}.\end{aligned} \tag{46}$$

I solve this problem with respect to λ. The first order condition is:

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

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

$$\lambda^* = \sqrt{\frac{e'\mathbb{E}(\mathcal{R}_i^{-1})e - e'\mathbb{E}(\mathcal{R}_i)^{-1}e}{\mathbb{E}(\mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i) - \mathbb{E}(\mathcal{Y}_i)'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i)}}. \tag{47}$$

Substituting (47) into (46) and applying the $1/T$ scaling factors to $\mathcal{R}_i$ and $\mathcal{Y}_i$ yield the expression of $\tilde{U}$ in Proposition 2.

The numerator and the denominator inside the square root of (47) are both weakly positive, and they are zero if and only if $\mathcal{R}_i^{-1}e$ and $\mathcal{R}_i^{-1}\mathcal{Y}_i$ are degenerate across individuals, respectively. To see why, let R and Y be matrices that have the same dimensions as

$\mathcal{R}_i$ and $\mathcal{Y}_i$. Define the functions:

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

Then the following result applied to $\mathcal{E}$ and $\mathcal{D}$ implies that the numerator and the denominator in (47) are weakly positive[10].

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

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

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

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

A.5 Proof of Theorem 2

In what follows, I show that (11) is the dual problem of (10), by rewriting the problem into a standard form of infinite-dimensional linear programming and applying the duality theorem. The same argument applies to (12).

I first rewrite (10) into a standard form of infinite-dimensional linear programming, for which I introduce additional notation. Recall that $\mathcal{M}_{W \times V}$ is a linear space of finite and countably 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 $\mathcal{W} \times \mathcal{V}$. Note that $\mathcal{F}_{W \times V}$ is a linear subspace of $\overline{\mathcal{F}}_{W \times V}$.

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 a linear space of finite and countably additive signed Borel measures on $\mathcal{W}$. Let $\overline{\mathcal{F}}_W$ be the dual space of $\mathcal{M}_W$, and let $\mathcal{F}_W$ be the space of all bounded Borel measurable functions on $\mathcal{W}$. Note that $\mathcal{F}_W$ 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 $g = (g_1, \dots, g_K, P_g)$ and $h = (\lambda_1, \dots, \lambda_K, f_h)$

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

be their generic elements. Then $\mathcal{H}$ is the dual space of $\mathcal{G}$. Define the dual pairing

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

Next, define a linear map $A : \mathcal{M}_{W\times V} \mapsto \mathcal{G}$ by

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

A is a bounded (hence continuous) linear operator because ϕ_ks are assumed to be bounded. Note that

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

It is straightforward to show:

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

Then:

$$\langle A(P), h\rangle = \sum_{k=1}^{K} \lambda_k \int \phi_k dP + \int f_h dP = \int \left[\sum_{k=1}^{K} \lambda_k \phi_k + f_h\right] dP \equiv \langle P, A^*(h)\rangle, \tag{48}$$

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

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

(48) shows that A^* is the adjoint of A. This is a key ingredient of the duality theorem.

With these notation, I rewrite (10) into a standard form of infinite-dimensional linear programming:

$$\min_{P\in\mathcal{M}_{W\times V}} \langle P, m\rangle \qquad \text{subject to} \qquad A(P) = c, \qquad P \geq 0, \tag{49}$$

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

With Assumption 5 and the continuity of A, the strong duality theorem holds (Anderson, 1983, Theorem 6), which implies that the optimal solution to (49) equals to the solution to:

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

which I can write more concretely as:

$$\max_{\lambda_1,\ldots,\lambda_K \in \mathbb{R},\, f_h \in \overline{\mathcal{F}}_W} \int f_h dP_W \qquad \text{subject to} \qquad \sum_{k=1}^{K} \lambda_k \phi_k + f_h \leq m. \tag{50}$$

Now I show that (50) simplifies to (11). I rearrange the constraint of (50):

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

The left-hand side does not involve v. Therefore:

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

Since (50) maximizes the integral of $f_h(w)$, the optimal f_h for fixed λs must satisfy:

$$f_h^*(w) = \min_{v \in \mathcal{V}} \left[m(w,v) - \sum_{k=1}^{K} \lambda_k \phi_k(w,v) \right] \tag{51}$$

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

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

with positive probability, one can increase the value of the objective by increasing f_h^* on a set of positive measure.

Substituting (51) into (50) then gives:

$$\max_{\lambda_1,\ldots,\lambda_K \in \mathbb{R}} \int \min_{v \in \mathcal{V}} \left[m(w,v) - \sum_{k=1}^{K} \lambda_k \phi_k(w,v) \right] dP_W(w).$$

The above display remains equivalent even if the signs of $(\lambda_1, \ldots, \lambda_K)$ are switched because λs are choice variables supported on $\mathbb{R}^K$. The problem then becomes:

$$\max_{\lambda_1,\ldots,\lambda_K \in \mathbb{R}} \int \min_{v \in \mathcal{V}} \left[m(w,v) + \sum_{k=1}^{K} \lambda_k \phi_k(w,v) \right] dP_W(w)$$

which is the expression in (11). □

A.6 Proof of Corollary 2

As in (50) in the proof of Theorem 2, the sharp lower bound of θ is given by

$$\max_{\lambda_1,\ldots,\lambda_K\in\mathbb{R},\ f_h\in\overline{\mathcal{F}}_W} \int f_h dP_W \qquad \text{subject to} \qquad \sum_{k=1}^{K} \lambda_k \phi_k + f_h \le m \tag{52}$$

where all notation follow the proof of Theorem 2. Similarly, the sharp upper bound of θ is given by

$$\min_{\lambda_1,\ldots,\lambda_K\in\mathbb{R},\ f_h\in\overline{\mathcal{F}}_W} \int f_h dP_W \qquad \text{subject to} \qquad \sum_{k=1}^{K} \lambda_k \phi_k + f_h \ge m. \tag{53}$$

Suppose that θ is point-identified but there is no such function $S^* \in \overline{\mathcal{F}}_W$ and real numbers $\lambda_1^*, \ldots, \lambda_K^* \in \mathbb{R}$ such that

$$\sum_{k=1}^{K} \lambda_k^* \phi_k + S^* = m$$

almost surely. Then the solution $(\lambda_1^l, \ldots, \lambda_K^l, S^l)$ to (52) satisfies

$$\sum_{k=1}^{K} \lambda_k^l \phi_k + S^l \le m$$

where inequality is strict on a set with positive Lebesgue measure on $\mathcal{W} \times \mathcal{V}$. Similarly, the solution $(\lambda_1^u, \ldots, \lambda_K^u, S^u)$ to (53) satisfies

$$\sum_{k=1}^{K} \lambda_k^u \phi_k + S^u \ge m$$

where inequality is strict on a set with positive Lebesgue measure on $\mathcal{W} \times \mathcal{V}$. Then:

$$\begin{aligned}
\mathbb{E}(S^l) &= \mathbb{E}\left(\sum_{k=1}^{K} \lambda_k^l \phi_k + S^l\right) \\
&< \mathbb{E}(m) \\
&< \mathbb{E}\left(\sum_{k=1}^{K} \lambda_k^u \phi_k + S^u\right) \\
&= \mathbb{E}(S^u)
\end{aligned}$$

where strict inequalities follow because the density of (W_i, V_i) is strictly positive. The above implies that the sharp lower bound $\mathbb{E}(S^l)$ is strictly less than the sharp upper bound $\mathbb{E}(S^u)$, which is contradiction since θ is assumed to be point-identified.

Conversely, suppose that there exists such $(S^*, \lambda_1^*, \ldots, \lambda_K^*)$. Then:

$$\mathbb{E}(S^*) = \mathbb{E}\left(S^* - \sum_{k=1}^{K} \lambda_k \phi_k\right) = \mathbb{E}(m) = \theta,$$

which proves that θ is point-identified by $\mathbb{E}(S^*)$. □

A.7 Proof of Proposition 3

It suffices to show that G_L 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, for any $t \in [0,1]$:

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

This is the definition of concavity. □

A.8 Proof of Proposition 4

When L exists and is finite, Proposition 3 implies that $L(\lambda)$ is also concave. Then $\hat{L}$ uniformly converges to L on any compact set $K \subseteq \mathbb{R}^K$, as in the proof of Theorem 2.7 in Newey and McFadden (1994):

$$\sup_{\lambda \in K} |\hat{L}(\lambda) - L(\lambda)| \xrightarrow{p} 0. \tag{54}$$

Let $\hat{\lambda} = \text{argmax}_\lambda\, \hat{L}(\lambda)$ and $\lambda_0 = \text{argmax}_\lambda\, L(\lambda)$. If there are multiple argmax's, choose any of them. Then for $\hat{\lambda}$ that is on a compact set $K \subseteq \mathbb{R}^K$:

$$\begin{aligned}
|L(\lambda_0) - \hat{L}(\hat{\lambda})| &\leq L(\lambda_0) - L(\hat{\lambda}) + |L(\hat{\lambda}) - \hat{L}(\hat{\lambda})| && \text{(triangle inequality)} \\
&= \hat{L}(\lambda_0) - L(\hat{\lambda}) + |L(\hat{\lambda}) - \hat{L}(\hat{\lambda})| + o_p(1) && \text{(by (54))} \\
&\leq \hat{L}(\hat{\lambda}) - L(\hat{\lambda}) + |L(\hat{\lambda}) - \hat{L}(\hat{\lambda})| + o_p(1) && (\hat{\lambda} \text{ is argmax}) \\
&\leq 2|L(\hat{\lambda}) - \hat{L}(\hat{\lambda})| + o_p(1) = o_p(1). && \text{(by (54))}
\end{aligned}$$

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, $L(\lambda)$ is a constant function, in which case consistency is immediate. If such ε exists, by Theorem 5.14 of Van der Vaart (2000):

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

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

A.9 Proof of Proposition 5

I can rewrite (22) as:

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

I can then replicate the argument of Theorem 2, obtaining (23) as the simplified dual of the above. □

A.10 Proof of Proposition 6

I can rewrite (25) as:

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

I can then replicate the argument of Theorem 2, obtaining (24) as the simplified dual of the above. □

B Identification under conditional moment restrictions

This section studies moment equality models that involve both conditional and unconditional moment restrictions. Consider the following extension of Assumption 4:

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

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

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

Under Assumption 6, I characterize the identified set of

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

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

To simplify notation, let A'_{ik} be the subvector of (W_i,V_i) that collects the variables not included in A_{ik}. I write any function $f(w,v)$ on $\mathcal{W}\times\mathcal{V}$ equivalently as $f(a_k,a'_k)$ on $\mathcal{A}_k\times\mathcal{A}'_k$, where $\mathcal{A}_k$ is the support of A_{ik} and $\mathcal{A}'_k$ is the support of A'_{ik}.

I assume the following regularity conditions:

Assumption 7. The following conditions hold.

(i) The distribution of (W_i,V_i) is absolutely continuous with respect to Lebesgue measure. In addition, its density p is L^∞ with respect to the Lebesgue measure.

(ii) $\mathcal{W} \times \mathcal{V}$ is a compact set in an Euclidean space.

(iii) $(m, \phi_1, \ldots, \phi_{K_U}, \psi_1, \ldots, \psi_{K_C})$ are L^∞ with respect to the Lebesgue measure.

(iv) The set

$$D = \Big\{ \Big(\int \phi_1 p \, d(w,v), \ldots, \int \phi_K p \, d(w,v), \\ \int \psi_1 p \, da'_k, \ldots, \int \psi_K p \, da'_k, \int mp \, d(w,v) \Big) \; \Big| \; p \in \mathcal{M}_{W \times V}, p \geq 0 \Big\}$$

is closed.

The first condition is restrictive, but it is useful enough for showing Proposition 1 based on it. The next three conditions of Assumption 7 are the counterparts of Assumption 5. A sufficient condition for Assumption 7 (iv) is that the joint density of (W_i, V_i) in the data generating process, or its observationally equivalent one, is strictly positive on $\mathcal{W} \times \mathcal{V}$ (Anderson, 1983, Theorem 9).

Under these assumptions, the following result characterizes the identified set I of θ.

Theorem 3. *Suppose Assumptions 6 and 7 hold. Let $\lambda_k \in \mathbb{R}$ for $k = 1, \ldots, K_U$, and let $\mu_k : \mathcal{A}_k \mapsto \mathbb{R}$ for $k = 1, \ldots, K_C$. Then $I = [L, U]$ where*

$$L = \max_{\{\lambda_k\}_{k=1}^{K_U}, \{\mu_k\}_{k=1}^{K_C}} \mathbb{E} \left[\min_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K_U} \lambda_k \phi_k(W_i, v) + \sum_{k=1}^{K_C} \mu_k(A_k(W_i, v)) \psi_k(W_i, v) \right\} \right] \tag{55}$$

and

$$U = \min_{\{\lambda_k\}_{k=1}^{K_U}, \{\mu_k\}_{k=1}^{K_C}} \mathbb{E} \left[\max_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K_U} \lambda_k \phi_k(W_i, v) + \sum_{k=1}^{K_C} \mu_k(A_k(W_i, v)) \psi_k(W_i, v) \right\} \right] \tag{56}$$

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

Proof. The proof focuses on showing (55). The same argument applies to (56).

Let $\mathcal{M}_{W \times V}$ to be the space of finite and countably additive signed Borel measures that are absolutely continuous with respect to the Lebesgue measure. Using absolute continuity, I identify an element of $\mathcal{M}_{W \times V}$ by its density $p : \mathcal{W} \times \mathcal{V} \mapsto \mathbb{R}$. Let p_W be the density of P_W. The identified set I is then 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 $\mathcal{A}_k'$. The second line represents unconditional moment restrictions and the third line represents conditional moment restrictions.

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

$$
\begin{aligned}
\min_{p \in \mathcal{M}_{W\times V},\, p \geq 0} & \int m(w,v)p(w,v)d(w,v) \qquad \text{subject to} \\
& \int \phi_k(w,v)p(w,v)d(w,v) = 0, \quad k = 1,\ldots,K_U, \\
& \int \psi_k(a_k,a_k')p(a_k,a_k')da_k' = 0, \text{ for all } a_k \in \mathcal{A}_k, \quad k = 1,\ldots,K_C, \\
& \int p(w,v)dv = p_W(w) \;\text{ for all } w \in \mathcal{W}.
\end{aligned}
\tag{57}
$$

Now I show that (55) is the dual of (57). I introduce additional notation. Let $L^2(\mathcal{W} \times \mathcal{V})$ be the space of all L^2 functions on $\mathcal{W} \times \mathcal{V}$, and let $L^2(\mathcal{W})$ be the space of all L^2 functions on $\mathcal{W}$. We also let $L^2(\mathcal{A}_k)$ be the space of all L^2 functions on $\mathcal{A}_k$.

Let $\mathcal{G}$ and $\mathcal{H}$ be $\mathcal{G} = \mathcal{H} = \mathbb{R}^K \times L^2(\mathcal{A}_1) \times \ldots \times L^2(\mathcal{A}_{K_C}) \times L^2(\mathcal{W})$. Denote their generic elements as $g = (g_1,\ldots,g_{K_U},\bar{g}_1,\ldots,\bar{g}_{K_C},f_g)$ and $h = (\lambda_1,\ldots,\lambda_{K_U},\mu_1,\ldots,\mu_{K_C},f_h)$, respectively. Note that $\mathcal{H}$ is a dual space of $\mathcal{G}$.

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

$$
A(p) = \left(\int \phi_1 p\, d(w,v), \ldots, \int \phi_K p\, d(w,v), \int \psi_k p\, da_1', \ldots, \int \psi_k p\, da_{K_C}', \int p\, dv \right).
$$

A is a bounded (hence continuous) linear operator because ϕ_ks and ψ_ks are assumed to be bounded. Define the dual pairing:

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

It is straightforward to show:

$$\iint \psi_k p \; da'_k \; \mu_k da_k = \int \psi_k \mu_k p \; d(w,v)$$

and

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

Then:

$$\langle A(P), h \rangle = \int \left[\sum_{k=1}^{K_U} \lambda_k \phi_k + \sum_{k=1}^{K_C} \mu_k \psi_k + f_h \right] p(w,v) d(w,v). \equiv \langle p, A^*(h) \rangle, \qquad (58)$$

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

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

(58) shows that A^* is the adjoint of A.

With Assumption 7 and the continuity of A, the strong duality theorem holds (Anderson, 1983, Theorem 6). The optimal solution to (57) equals to the solution to:

$$\max_{\lambda_1,\ldots,\lambda_{K_U},\mu_1,\ldots,\mu_{K_C},f_h} \int f_h(w) p_w(w) dw \qquad \text{subject to} \qquad \sum_{k=1}^{K_U} \lambda_k \phi_k + \sum_{k=1}^{K_C} \mu_k \psi_k + f_h \leq m. \quad (59)$$

Simplifying (59) as in the proof of Theorem 2 yields the expression in (55).

□

Proof of Theorem 3 has the following implication about point-identification of θ, which generalizes Corollary 2.

Corollary 3. *Suppose that assumptions of Theorem 3 hold. Suppose also that the joint density of (W_i, V_i) is strictly positive on $\mathcal{W} \times \mathcal{V}$. Then θ is point-identified if and only if there exists a function S^* which is a linear functional on the projection of $\mathcal{M}_{W \times V}$ onto $\mathcal{W}$, real numbers $\lambda_1^*, \ldots, \lambda_K^* \in \mathbb{R}$ and functions $\mu_1^*, \ldots, \mu_K^*$ which are $L^2(\mathcal{A}_1), \ldots, L^2(\mathcal{A}_{K_C})$ functions, respectively, such that:*

$$m(W_i, v) + \sum_{k=1}^{K_U} \lambda_k \phi_k(W_i, v) + \sum_{k=1}^{K_C} \mu_k(A_k(W_i, v)) \psi_k(W_i, v) = S^*(W_i)$$

almost surely on $\mathcal{W} \times \mathcal{V}$. When such S^* exists, θ is identified by $\theta = \mathbb{E}(S^*(W_i))$.

Proof. As in (59) during the proof of Theorem 3, the sharp lower bound of θ is given by

$$\max_{\lambda_1,\ldots,\lambda_{K_U},\mu_1,\ldots,\mu_{K_C},f_h} \int f_h(w)p_w(w)dw \qquad \text{subject to} \qquad \sum_{k=1}^{K_U} \lambda_k \phi_k + \sum_{k=1}^{K_C} \mu_k \psi_k + f_h \leq m.$$

where all notation follow the proof of Theorem 3. Similarly, the sharp upper bound of θ is given by

$$\min_{\lambda_1,\ldots,\lambda_{K_U},\mu_1,\ldots,\mu_{K_C},f_h} \int f_h(w)p_w(w)dw \qquad \text{subject to} \qquad \sum_{k=1}^{K_U} \lambda_k \phi_k + \sum_{k=1}^{K_C} \mu_k \psi_k + f_h \geq m.$$

Corollary 3 can then be proved by replicating the proof of Corollary 2. □

C Supplementary materials for application

C.1 Results without removing observations or de-noising

This subsection presents estimation results in Section 8.5 without removing individuals with smallest variations in their reported incomes or without the de-noising step. The results will show that that estimation results do not change qualitatively by removing observations or de-noising the data.

Tables 6 and 7 present confidence intervals for $\mathbb{E}(\rho_i)$, $\text{Var}(\rho_i)$ and the CDF of ρ_i without removing observations. The confidence intervals do not qualitatively differ to Tables 4 and 5 in the main text, in the sense that upper confidence limit of $\mathbb{E}(\rho_i)$ is significantly less than 1, lower confidence limit of $\text{Var}(\rho_i)$ is strictly positive and the CDF of ρ_i has confidence limits away from 0 and 1.

C.2 Simulation based de-noising

Before describing how I separated ε_{it} from Y_{it}, I first describe the simulation based de-noising method proposed in Arellano and Bonhomme (2018). They considered a model

$$Z = X + \varepsilon$$

where all variables are scalar and Z is observed but X and ε are not. Instead, distribution of ε is known. A researcher's objective is to obtain the distribution of X, or the data drawn from it.

The idea of Arellano and Bonhomme (2018) is then to find a distribution of X that minimizes the second order Wasserstein distance (W_2) between Z and $X + \varepsilon$. It is known that W_2 metrizes convergence in the second mean, which means that a zero W_2 between two distributions implies that they are equal in distribution. They proposed a procedure for obtaining i.i.d. draws of X given the observed data of Z and the simulated i.i.d. draws of ε.

I use their idea to separate transitory income from the observed income data, obtaining simulated draws of permanent income. I assume that the transitory income process in the application, ε_{it}, follows i.i.d. zero-mean Normal distribution whose variance equals to the variance estimate of the transitory income in Guvenen (2009). I then simulate $K = 200$ i.i.d. draws of $\varepsilon_k = (\varepsilon_{k1}, \ldots, \varepsilon_{kT})'$. Then, given the permanent income data $\tilde{Y}_j = (\tilde{Y}_{j1}, \ldots, \tilde{Y}_{jT})'$, $j = 1, \ldots, N$, I obtain the *simulated* income data $\hat{Y}_{jk}$ defined by:

$$\hat{Y}_{jk} = \tilde{Y}_j + \varepsilon_k,$$

giving the simulated income data of size NK.

I compare this to the observed income data $Y_i = (Y_{i1}, \ldots, Y_{iT})'$, $i = 1, \ldots, N$. I compute the distance between the simulated and the observed data using the second order Wasserstein distance:

$$\begin{aligned} W_2^2(\{Y_i\}, \{\hat{Y}_{ij}\}) = &\min_{0 \le p_{ijk} \le 1} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{K} p_{ijk} ||Y_i - \hat{Y}_{jk}||^2 \\ &\text{subject to} \quad \sum_{i=1}^{N} p_{ijk} = 1, \quad \sum_{j=1}^{N} \sum_{k=1}^{K} p_{ijk} = 1. \end{aligned}$$

I then obtain the simulated permanent income $\tilde{Y}_j = (\tilde{Y}_{j1}, \ldots, \tilde{Y}_{jT})'$ by solving:

$$\operatorname*{argmin}_{\tilde{Y}_j} W_2^2(\{Y_i\}, \{\hat{Y}_{ij}\}),$$

giving a dataset with $N = 800$ individuals and $T = 15$ waves.

Estimation results in Section 8.5 in the main text are not affected qualitatively by this de-noising procedure. Tables 8 and 9 present estimation results where I did not apply the de-noising step. Estimation results with and without de-noising are qualitatively similar. Upper confidence limit of $\mathbb{E}(\rho_i)$ is significantly less than 1, lower confidence limit of $\text{Var}(\rho_i)$ is strictly positive and the CDF of ρ_i has confidence limits away from 0 and 1.

D Tables

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

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

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

Table 2: Coverage probabilities of the inference procedure with sample size of $N = 750$. Nominal coverage probability is 0.9.

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

Table 3: Coverage probabilities of the inference procedures with sample size of $N = 1000$. Nominal coverage probability is 0.9.

	Confidence interval of $\mathbb{E}(\rho_i)$	Confidence interval of $\mathrm{Var}(\rho_i)$
First model	[0.456, 0.615]	[0.067, 0.292]
Second model	[0.264, 0.583]	[0.000, 0.701]

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

$\mathbb{P}(\rho_i \leq r)$	First model	Second model
$r = 0.0$	[0.000, 0.362]	[0.000, 0.752]
$r = 0.1$	[0.005, 0.428]	[0.004, 0.800]
$r = 0.2$	[0.025, 0.548]	[0.099, 0.818]
$r = 0.3$	[0.066, 0.627]	[0.085, 0.834]
$r = 0.4$	[0.104, 0.713]	[0.135, 0.879]
$r = 0.5$	[0.153, 0.848]	[0.205, 0.914]
$r = 0.6$	[0.209, 0.895]	[0.200, 0.983]
$r = 0.7$	[0.290, 0.944]	[0.286, 0.986]
$r = 0.8$	[0.372, 0.975]	[0.319, 1.000]
$r = 0.9$	[0.471, 0.994]	[0.353, 1.000]
$r = 1.0$	[0.550, 1.000]	[0.427, 1.000]

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

	Confidence interval of $\mathbb{E}(\rho_i)$	Confidence interval of $\mathrm{Var}(\rho_i)$
First model	[0.415, 0.652]	[0.073, 0.235]
Second model	[0.262, 0.692]	[0.000, 0.659]

Table 6: Confidence interval for $\mathbb{E}(\rho_i)$ and $\mathrm{Var}(\rho_i)$, for the dataset of $N = 800$ that does not remove individuals with small variation in their reported incomes and with additional moment restrictions. Nominal coverage probability is 0.9.

$\mathbb{P}(\rho_i \leq r)$	First model	Second model
$r = 0.0$	[0.000, 0.422]	[0.000, 0.665]
$r = 0.1$	[0.001, 0.482]	[0.014, 0.766]
$r = 0.2$	[0.024, 0.613]	[0.075, 0.804]
$r = 0.3$	[0.054, 0.674]	[0.104, 0.827]
$r = 0.4$	[0.090, 0.770]	[0.204, 0.872]
$r = 0.5$	[0.125, 0.845]	[0.217, 0.935]
$r = 0.6$	[0.188, 0.930]	[0.208, 0.979]
$r = 0.7$	[0.256, 0.959]	[0.250, 1.000]
$r = 0.8$	[0.330, 0.987]	[0.310, 1.000]
$r = 0.9$	[0.411, 0.998]	[0.369, 1.000]
$r = 1.0$	[0.498, 1.000]	[0.428, 1.000]

Table 7: Confidence intervals for $\mathbb{P}(\rho_i \leq r)$, for the dataset of $N = 800$ that does not remove individuals with small variation in their reported incomes. Nominal coverage probability is 0.9.

	Confidence interval of $\mathbb{E}(\rho_i)$	Confidence interval of $\mathrm{Var}(\rho_i)$
First model	[0.451, 0.615]	[0.050, 0.293]
Second model	[0.242, 0.633]	[0.000, 0.700]

Table 8: Confidence interval for $\mathbb{E}(\rho_i)$ and $\mathrm{Var}(\rho_i)$, for the dataset without the de-noising step of size $N = 760$ with removing individuals with small variation in their reported incomes. Nominal coverage probability is 0.9.

$\mathbb{P}(\rho_i \leq r)$	First model	Second model
$r = 0.0$	[0.000, 0.364]	[0.000, 0.715]
$r = 0.1$	[0.012, 0.413]	[0.006, 0.765]
$r = 0.2$	[0.026, 0.510]	[0.037, 0.845]
$r = 0.3$	[0.084, 0.584]	[0.122, 0.843]
$r = 0.4$	[0.118, 0.725]	[0.152, 0.882]
$r = 0.5$	[0.150, 0.826]	[0.170, 0.936]
$r = 0.6$	[0.232, 0.879]	[0.216, 0.983]
$r = 0.7$	[0.309, 0.934]	[0.283, 1.000]
$r = 0.8$	[0.393, 0.982]	[0.324, 1.000]
$r = 0.9$	[0.493, 0.990]	[0.359, 1.000]
$r = 1.0$	[0.558, 1.000]	[0.384, 1.000]

Table 9: Confidence intervals for $\mathbb{P}(\rho_i \leq r)$, for the dataset without the de-noising step of size $N = 760$ with removing individuals with small variation in their reported incomes . Nominal coverage probability is 0.9.