UNIVERSITY OF TECHNOLOGY SYDNEY

UTS Business School –
Department of Economics
Dr Chau Chak Wing Building
14 – 28 Ultimo Road, Ultimo NSW 2007

T: +61 2 9514 7713
economics@uts.edu.au

Department of Economics

Working Paper Series

'Identification and estimation of dynamic random coefficient models'

Wooyong Lee [1]

[1] The University of Technology Sydney

No. 03 / 2022 **October 2022**

Identification and estimation of dynamic random coefficient models

Wooyong Lee[*]

October 27, 2022

Abstract

I study panel data linear models with predetermined regressors (e.g. lagged dependent variables) that allow the coefficients as well as the intercept to be individual-specific, permitting unobserved heterogeneity in the effects of regressors on the dependent variable. I show that the model is not point-identified in a short panel context but rather partially identified, and I characterize sharp identified sets of the mean, variance, and CDF of the coefficient distributions. The characterization is general, allowing discrete, continuous, and unbounded data. A computationally efficient estimation and inference procedure is proposed, based on a fast and precise 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. The results suggest substantial unobserved heterogeneity in earnings persistence, which implies that households face different levels of earnings risk that lead to heterogeneity in their consumption and savings behaviors.

Keywords: panel data, heterogeneous effects, partial identification.

[*] Economics Discipline Group, UTS Business School, University of Technology Sydney. Address: 14-28 Ultimo Road, Ultimo, NSW 2007, Australia. Tel: (02) 9514 3074. Email: wooyong.lee.econ@gmail.com

1 Introduction

Panel data linear models with predetermined regressors (e.g. lagged dependent variables) are extremely popular in empirical research (Arellano and Bond, 1991; Blundell and Bond, 1998). Most of these models allow for fixed effects, which are individual-specific intercepts that permit unobserved heterogeneity in the levels of dependent variables. Fixed effects offer a flexible method of controlling for unobserved heterogeneity in those levels, helping researchers to explore diverse research questions, such as the effectiveness of a public policy. Fixed effects models are well understood for short panel data (i.e., panel data with a small number of waves).

In addition to heterogeneity in the levels of dependent variables, there is ample evidence that individuals show unobserved heterogeneity in the effects of regressors on dependent variables. For example, firms demonstrate different degrees of labor efficiency for production; individuals receive different rates of return on education, and households show different degrees of persistence in their earnings with respect to their past earnings. Such heterogeneous effects are crucial mechanisms for heterogeneous responses to exogenous shocks and policies, such as employment subsidies, tuition subsidies, and income tax reform. Heterogeneous effects also have a first-order influence on the outcomes of various economic models. For example, heterogeneity in earnings persistence governs heterogeneity in the earnings risk that households experience, which leads to heterogeneous motives for precautionary savings in the lifecycle model of consumption.

This paper studies a panel data linear model with predetermined regressors that allows for unobserved heterogeneity in both the effects of regressors and the levels (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 potentially correlated with its future values. In this model, both the coefficient (β_{i1}) and the intercept (β_{i0}) are individual-specific, reflecting heterogeneity in the effects of regressors and the levels, respectively. The model also allows the lagged dependent variable $Y_{i,t-1}$ 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. An important work by Chamberlain (1993), recently published as Chamberlain (2022), showed that the mean of β_is in dynamic random coefficient models is not point-identified, implying that it is not consistently estimable. Since this negative result in the 1990s, there has been little progress in the literature. Arellano and Bonhomme (2012) showed that when regressors are binary, the mean of β_is for some subpopulations is identifiable and hence consistently estimable; however, they did not provide a general identification result that allows for non-binary regressors. Most research on random coefficient models in short panels focuses on non-dynamic contexts (Chamberlain, 1992; Wooldridge, 2005; Arellano and Bonhomme, 2012; Graham and Powell, 2012), but such contexts exclude important dynamic mechanisms, such as feedback from the current dependent variable to the future regressors. For example, a firm's labor purchase decision next year might depend on this year's output, because the firm might learn about its own labor efficiency from that output. Moreover, a researcher might also be interested in the dynamic mechanisms. For example, a household's earnings persistence with respect to its past earnings is an important parameter because high earnings persistence increases the duration of earnings shocks, reducing a household's consumption smoothing ability, to the detriment of household welfare.

This paper is the first to present a general identification result for dynamic random coefficient models in a short panel context. 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 estimation and inference procedure for these features, an essential step of which is to employ a fast and precise algorithm to solve global polynomial optimization problems. The procedure is then applied with the aim of learning about unobserved heterogeneity in lifecycle earnings dynamics across U.S. households in the Panel Study of Income Dynamics (PSID) dataset. These are presented in three steps.

First, I show that dynamic random coefficient models are partially identified, which means that there exist finite lower and upper bounds for the parameters of interest, such as the mean, variance, and CDF of β_is. The result is general, allowing the data and coefficients to be discrete, continuous, or unbounded. I provide a simple expression for the bounds of the mean of β_is, which explicitly shows that the bounds are finite even if the data and coefficients are unbounded, as long as certain moments of data are finite. These results are obtained by recasting the identification problem into a linear programming problem (Honoré and Tamer, 2006; Mogstad, Santos, and Torgovitsky, 2018; Torgovitsky, 2019), which becomes an infinite-dimensional problem when the 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 the parameters of interest.

Second, I show that the sharp bounds can be computed efficiently by exploiting the linear structure of the model. Computing sharp bounds obtained from the 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. I show that, for random coefficient models, the inner objective function is a polynomial. I then use the semidefinite relaxation algorithm (Lasserre, 2010, 2015), a fast and precise algorithm for solving the global polynomial optimization problem, to efficiently compute the sharp bounds. The algorithm also delivers a computationally tractable inference about the parameters, based on testing moment inequalities (Chernozhukov, Lee, and Rosen, 2013; Romano, Shaikh, and Wolf, 2014; Chernozhukov, Chetverikov, and Kato, 2019; Bai, Santos, and Shaikh, 2022). 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 processes are key inputs 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. I investigate unobserved heterogeneity in the earnings of U.S. households in the Panel Study of Income Dynamics (PSID) dataset. Guvenen (2007, 2009) pointed out that, when allowing for unobserved heterogeneity in the time trend of earnings (known as a heterogeneous income profile, HIP), the earnings persistence of the permanent income process is estimated to be significantly smaller than 1, with the latter being the estimate from the model without the heterogeneity in the time trend (known as a restricted income profile, RIP). I estimate a more general model that also permits unobserved heterogeneity in the earnings persistence itself, and I find that HIP and RIP yield similar estimates of the average earnings persistence that are both significantly smaller than 1. This suggests that misspecifying HIP as RIP, or vice versa, may not lead to serious misspecification when earnings persistence is allowed to be heterogeneous. I also find evidence of substantial unobserved heterogeneity in the earnings persistence itself, which implies that households face different levels of earnings risk, leading to heterogeneity in their consumption and savings behaviors.

Identification results from this paper can be extended generally to other structural models to allow for heterogeneous effects. For example, the results can be applied to allow for individual-specific coefficients and intercepts in probit and logit regressions. They can also be applied to vector-valued regressions, such as panel data Vector Autoregressive (VAR) models and the systems of panel data regressions.

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

The remainder of this paper is structured as follows. Section 2 introduces a dynamic random coefficient model. Sections 3 and 4 present identification results about the model. Sections 5 and 6 introduce estimation, inference, and computation methods. Section 7 reviews the performance of the inference method through simulation. Section 8 applies the methods to lifecycle earnings dynamics. Section 9 concludes. All proofs and tables appear in the Appendix.

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:

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

so that the error term is mean-independent of the full history of $\{Z_{it}\}$ (strict exogeneity) and of the current history of $\{X_{it}\}$ (sequential exogeneity). The presence of a sequentially exogenous regressor $\{X_{it}\}$ makes (1) a dynamic model. For example, the lagged dependent variable $Y_{i,t-1}$ can be included in X_{it}.

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 coefficients (γ_i, β_i) are unobserved random variables that follow a nonparametric distribution, and they can freely correlate among themselves and with (Z_i, X_{i1}). This is how a random coefficient model extends a fixed effects model.

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

I consider a 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. Identification 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 allows for computationally feasible estimation and inference. 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 case m is set to be $m = \mathbf{1}(\beta_i \leq b)$, which yields $\theta = \mathbb{E}(\mathbf{1}(\beta_i \leq b)) = \mathbb{P}(\beta_i \leq b)$.

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

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

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, where Y_{it} is the log-earnings net of demographic variables, an important input in the lifecycle model of consumption and savings behavior[2]. The earnings persistence, β_i, governs the earnings risk experienced by households, which is a fundamental motive for precautionary savings. Specifying an earnings process that highlights features of real data is important for drawing conclusions from the model of consumption and savings behavior. The literature often 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 the time trend is heterogeneous. 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 at close to 1 (e.g. Abowd and Card, 1989; Topel and Ward, 1992). I find later in Section 8 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. 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.

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

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

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

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 this model, Y_{it} may be taken as strictly exogenous, meaning that the future earnings stream is unaffected by 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 is γ_{i1}, 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, and hence avoid detriment to household welfare. As with Example 1, the literature focuses on models with no coefficient heterogeneity[3].

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 remaining agnostic about the evolution of assets over time (i.e., under nonparametric evolution of the assets).

The results from this paper also encompass a multivariate version of (1), a system of random coefficient models. For example, a researcher can combine (3) and (4) and consider a joint lifecycle model of earnings and consumption behavior. The resulting model allows the coefficients from the two processes to freely correlate among themselves and with (Y_{i0}, A_{i1}), allowing for correlation between earnings and consumption processes. Full description of the multivariate model can be found in the Online Appendix B.1.

3 Identification of the mean

This section and the following section present identification results for the dynamic random coefficient model defined in (1). This section focuses on identification of the mean of random coefficients, which provides intuition for the general result in the next section.

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 mean of the first entry of β_i.

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

In the following subsections, I show that μ_e is generally not point-identified (Section 3.1). I then show that μ_e is non-trivially partially identified (Section 3.2). The results of this section are special cases of the general results in Section 4 and the Online Appendix B.2.

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:

$$Y_{it} = \gamma_i + \beta_i Y_{i,t-1} + \varepsilon_{it}, \qquad \mathbb{E}(\varepsilon_{it}|\gamma_i, \beta_i, Y_i^{t-1}) = 0, \qquad t = 1, 2. \tag{5}$$

The following proposition states that $\mathbb{E}(\beta_i)$ is not point-identified in this model, which implies that there is no consistent estimator for $\mathbb{E}(\beta_i)$. This proposition is an application of the general result in the Online Appendix B.2.

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, under the regularity conditions stated as Assumption 7 in the Online Appendix, $\mathbb{E}(\beta_i)$ is not point-identified.*

Chamberlain (1993), recently published as Chamberlain (2022), 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 complements this result, showing that point identification also fails with stronger assumptions and continuous data. Failure of point identification in both the discrete and continuous cases in (5) suggests that this is a general feature of dynamic random coefficient models.

Proof of Proposition 1 uses that $\mathbb{E}(\beta_i)$ is point-identified if and only if there exists an unbiased estimator of β_i in individual time series, something which is worth stating separately:

Lemma 1. *Under the assumptions of Proposition 1, $\mathbb{E}(\beta_i)$ is point-identified if and only if there exists 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 $\mathbb{E}(\beta_i) = \mathbb{E}(S^*(Y_{i0}, Y_{i1}, Y_{i2}))$.

I prove Proposition 1 by showing that there is no unbiased estimator of β_i. The intuition for Lemma 1 is as follows. Since the distribution of β_i is unrestricted, information on individual β_i can be obtained only from its individual time series. In a long panel context, such information can be obtained by a time series estimator of β_i that is consistent as $T \to \infty$. In a short panel context, however, such an estimator is not reliable because T is finite. Lemma 1 shows that a time series estimator that is unbiased for finite T is the only reliable information on β_i in short panels when it comes to point identification.

3.2 Partial identification

A natural question following the last subsection is whether the data are at all informative about $\mu_e = \mathbb{E}(e'V_i)$, or whether they provide no information. This subsection shows that the data are informative about μ_e. I show that there are finite bounds L and U such that

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

which are estimable with data.

I first 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

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

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; Andrews and Shi, 2013). I choose the class of g to be the set of polynomial functions and use its finite subset for estimation and inference. Such a finite subset contains less information than (7), but it yields a computationally feasible estimation and inference procedure. Partial identification results based on the conditional moment restriction of (7) are established in the Online Appendix B.2.

Consider the following assumptions:

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

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

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

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

Assumption 1 states that the dynamic random coefficient model is correctly specified. Assumption 2 is a no-multicollinearity assumption imposed on individual time series. This is stronger than the assumption that $\mathbb{E}(\sum_{t=1}^T R_{it}R_{it}')$ 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 be learned only from its individual data[4]. Assumption 3 considers a specific choice of unconditional moment restrictions that are implications of (7). The first restriction states that the "error term" (ε_{it}) is orthogonal to the "explained term" ($R_{it}'V_i$). The second 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 is absolutely continuous with respect to the Lebesgue measure. For brevity of notation, define*

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

Then $L \leq \mu_e \leq U$ where

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

and

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

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. In addition, $[L, U]$ are the sharp bounds of μ_e if Assumption 3 is replaced

[4] Graham and Powell (2012) studied violation of Assumption 2 in a non-dynamic context.

by the following implication of Assumption 3:

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

The closed-form expressions in Theorem 1 provide intuition on when L and U are finite. In particular, L and U are finite even if (Y_i, R_i, V_i) are unbounded, as long as $\mathbb{E}(\mathcal{R}_i)$, $\mathbb{E}(\mathcal{R}_i^{-1})$, $\mathbb{E}(\mathcal{Y}_i)$, $\mathbb{E}(\mathcal{R}_i^{-1}\mathcal{Y}_i)$, and $\mathbb{E}(\mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i)$ are finite. Note that $\mathcal{R}_i$ is the squared design matrix of individual time series, and that $\mathcal{R}_i^{-1}\mathcal{Y}_i$ is the OLS estimator of V_i from individual time series.

I now explain the intuition behind Theorem 1, focusing on the upper bound U. Consider a Lagrangian where the objective function is the parameter of interest $e'V_i$ and the constraints are the moment functions in (9):

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

where $\lambda \in \mathbb{R}$ and μ has the same dimension as R_{it}. Note that $\mathbb{E}(Q) = \mathbb{E}(e'V_i) = \mu_e$ because the constraints have zero expectations.

If I substitute $\varepsilon_{it} = Y_{it} - R_{it}V_i$ into Q and use the notation of $\mathcal{R}_i$ and $\mathcal{Y}_i$ in Theorem 1, I obtain the expression:

$$Q(\lambda, \mu, W_i, V_i) = e'V_i + \lambda\mathcal{Y}_i'V_i - \lambda V_i'\mathcal{R}_iV_i + \mu'\mathcal{Y}_i - \mu'\mathcal{R}_iV_i.$$

This is a quadratic polynomial in V_i whose second-order derivative is

$$\frac{d^2Q}{dV_i dV_i'} = -2\lambda\mathcal{R}_i.$$

If $\lambda > 0$, 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." Then:

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

which implies:

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

This shows that $\mathbb{E}(P)$ is an upper bound of μ_e for any $\lambda > 0$ and μ. I then obtain the smallest upper bound by minimizing $\mathbb{E}(P)$ with respect to $\lambda > 0$ and μ:

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

This coincides with U in Theorem 1, which is the sharp upper bound of μ_e under (9). The lower bound can be obtained by repeating the same process with $\lambda < 0$.

4 Identification of the general parameters

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 $\phi_k : \mathcal{W} \times \mathcal{V} \mapsto \mathbb{R}$ are moment functions and $K \in \mathbb{N}$ is the number of moments.

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 generic moment equalities without connection to random coefficient models. A more general case that also involves conditional moment restrictions is studied in the Online Appendix B.2.

The following example illustrates how the moment restrictions considered in the previous section are a special case of Assumption 4.

Example 3. 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

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

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

I characterize the identified set of θ under Assumption 4 and additional regularity conditions stated later. To do so, I first recast the identification problem into a linear programming problem. I then show that its dual representation yields a tractable characterization of the identified set.

Let $P_{W,V} \in \mathcal{M}_{W \times V}$, where $\mathcal{M}_{W \times V}$ is the linear space of finite and countably additive signed Borel measures on $\mathcal{W} \times \mathcal{V}$, 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:

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

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. Dependence of I on m, P_W, and the ϕ_ks are suppressed in the notation.

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

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

The constraint $\int dP = 1$ is omitted because it is implied by the last line of (10). Note that P_W is a probability distribution.

(10) is a linear program 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 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 dependent variables and V_i contains all random coefficients. For the random coefficient model with R regressors and T waves, P is a distribution on an $(RT + R + T)$-dimensional space.

My approach is to use the dual representation of (10) obtained by the duality theorem for infinite-dimensional linear programming (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 a Euclidean space.

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

(iii) The following set is closed:

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

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 distribution, 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

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

θ is point-identified if and only if $L = U$. Proof of Theorem 2 then implies a necessary and sufficient condition for point identification of θ, something which is worth stating separately:

Lemma 2. *Suppose that the 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 $\theta = \mathbb{E}(S^*(W_i))$.

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

Theorem 2 and Lemma 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 (Schennach, 2014; Li, 2018). In general, it is not obvious that Theorem 2 leads to a computationally feasible estimation and inference procedure. I show in the next sections that, for dynamic random coefficient models, I can obtain a computationally tractable estimation and inference procedure by exploiting the fact that it is a linear model.

5 Estimation and inference

This section explains the 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, instead observing a finite sample $(W_1, \ldots, W_N)$ of size N which is 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), which is equivalent to considering the empirical version of (10) where P_W is replaced by the empirical distribution $\hat{P}_W$. I 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.

The plug-in bound is used as a key object for estimation and inference, but it is not straightforward to compute. Its computation involves solving two types of optimization problem: the inner optimization problem over $\mathcal{V}$ and the outer optimization problem with respect to $\lambda_1, \ldots, \lambda_K$. Each problem presents its own difficulties. The inner problem must be solved globally, but its objective function is not necessarily convex. It must also be solved quickly because it must be solved for each i and each step of the outer problem. The outer problem must be solved globally, too, and it might be an optimization over a large dimensional space. The next section discusses how to tackle these computational issues. In this section, I discuss estimation and inference, assuming that the two optimization problems can be solved numerically.

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

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 the 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 through replication of the analysis of M-estimation. The regularity conditions of M-estimation are satisfied thanks to G_L being concave in λ.

Proposition 2. *$G_L(\lambda, W_i)$ is globally concave in λ, which implies global concavity of $L(\lambda)$.*

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

5.2 Inference

This subsection discusses construction of a confidence interval for the parameter $\theta \in \mathbb{R}$ whose sharp identified set is $[L, U]$ in Theorem 2. Any value $\theta \in [L, U]$ must satisfy:

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

which implies

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

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

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

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

The degree of conservatism in (18) relative to (17) depends on how much information is contained in (18) relative to (17). While a formal analysis of a comparison of (17) and (18) is beyond the scope of this paper[5], 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 $\lambda_L^* = \operatorname{argmax}_\lambda \mathbb{E}(G_L(\lambda, W_i))$ and $\lambda_U^* = \operatorname{argmin}_\lambda \mathbb{E}(G_U(\lambda, W_i))$, and the inequalities are loose for λs that are distant from them (because G_L is concave and G_U is convex). This

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

means that most of the information in (17) is contained in the neighborhood of λ_L^* and λ_U^*.[6] Second, the concavity of G_L (and convexity of G_U) implies that G_L and G_U are continuous, which means that consideration of a finite set of points in the neighborhood of λ_L^* and λ_U^* does not lead to serious loss of information, compared with consideration of 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 their neighborhoods. I review performance of this strategy through simulation in Section 7.

Note that (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; Chernozhukov, Chetverikov, and Kato, 2019; Bai, Santos, and Shaikh, 2022) proposes procedures for computing a confidence interval $[L_\alpha, U_\alpha]$ of θ that has an asymptotic size of α for large Λ_F. Among the proposed methods, a procedure based on multiplier bootstrap by Chernozhukov, Chetverikov, and Kato (2019) is particularly appealing, because the bootstrap does not require re-computation of G_L and G_U, which have high computational costs. Their procedure employs the following test statistic, computed for each $\theta \in \mathbb{R}$:

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

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

T_{CCK} is then compared with 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. It then follows that the confidence interval is the set of θ for which $T_{CCK}(\theta) \leq c_{CCK}(\alpha)$. Note that $c_{CCK}(\alpha)$ does not depend on θ, because c_{CCK} does not depend on it.

[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.

This allows for a computationally efficient search of θ, since $c_{CCK}(\alpha)$ can be computed only once and be fixed.

The inference procedure naturally extends to a vector-valued parameter $\theta \in \mathbb{R}^d$, through consideration of (17) for every entry of θ. For example, the moment inequalities for $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$ are:

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

where G_{Uk} and G_{Lk} are G_L and G_U in (17) for θ_k, $k = 1, 2$. Inference can then be performed via the same procedure, yielding 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 second moments.

5.3 Estimation and inference under over-identification

In practice, the plug-in bound may yield an empty set, in which case $\hat{L}$ diverges to $+\infty$ and $\hat{U}$ diverges to $-\infty$. This happens when the empirical distribution $\hat{P}_W$ does not satisfy the moment restrictions, which may occur even if the population distribution P_W satisfies the restrictions. In this case, the empirical version of (10) (where P_W is replaced with $\hat{P}_W$) does not have a feasible solution, resulting in an empty plug-in bound. This scenario is comparable with over-identification in the generalized method of moments (GMM) estimation, where the GMM objective may be strictly positive in the sample even if the moments are correctly specified.

There are two approaches for addressing this issue. First, a researcher may obtain a point estimate that minimizes the distance between the model and the data. Second, the researcher may directly obtain a confidence interval without insisting on a point estimate, assuming that the model is correctly specified. This subsection discusses these two approaches, summarizing the full discussion in the Online Appendix B.3.

For the first approach, consider the following relaxation of the moment restrictions:

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

where $\delta \geq 0$, which reduces to Assumption 4 when $\delta = 0$. This can be regarded as an absolute-value GMM criterion. Then, it can be shown that the smallest δ that allows (20) to hold with the empirical distribution, denoted by δ^*, is given by:

$$\max_{\lambda_1,\ldots,\lambda_K} \frac{1}{N}\sum_{i=1}^{N} \min_{v\in\mathcal{V}} \left\{ \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} \qquad \text{subject to} \qquad \sum_{k=1}^{K} |\lambda_k| \leq 1, \tag{21}$$

which can be computed using the computation methods in the next section. It can also be shown that, for $\delta \geq \delta^*$, the plug-in lower bound $\tilde{L}$ under the relaxation (20) is given by the bound with a negative L^1 penalty:

$$\tilde{L} = \max_{\lambda_1,\ldots,\lambda_K} \left[\frac{1}{N}\sum_{i=1}^{N} \min_{v\in\mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} - \delta \sum_{k=1}^{K} |\lambda_k| \right]. \tag{22}$$

The relaxed upper bound $\tilde{U}$ is given similarly to (22), with a positive L^1 penalty. When $\delta > \delta^*$, $\tilde{L}$ (and $\tilde{U}$) computes the smallest (and largest) value of θ among those that attain the *near-minimum* of the absolute-value GMM criterion in (20).

Although (22) resolves the empty set problem, it has two drawbacks. First, it is an ad hoc approach, with no formal justification for why the relaxation of moment conditions is a constructive idea. Second, the procedure may yield a point-identified set (or a small interval) even if the model is partially identified. While the literature deals with the second problem by choosing δ that is substantially larger than δ^* (Mogstad, Santos, and Torgovitsky, 2018), the question of how much larger it should be remains unresolved. In the rest of this subsection, I discuss a more principled approach, which is to directly compute a confidence interval without insisting on a point estimate.

Note that the 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 the plug-in bound problems. The inference procedure is valid regardless of whether the plug-in bound is empty; the issue here 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 to consider a grid of positive real numbers $\{\delta_1, \ldots, \delta_M\}$ such that $\delta_m > \delta^*$ for all $m \in \{1, \ldots, M\}$. Then, for each δ_m, I compute solutions to the relaxed plug-in bounds:

$$\tilde{\lambda}_L(\delta_m) = \operatorname*{argmax}_{\lambda_1,\ldots,\lambda_K} \left[\frac{1}{N}\sum_{i=1}^{N} \min_{v\in\mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} - \delta_m \sum_{k=1}^{K} |\lambda_k| \right],$$

$$\tilde{\lambda}_U(\delta_m) = \operatorname*{argmin}_{\lambda_1,\ldots,\lambda_K} \left[\frac{1}{N}\sum_{i=1}^{N} \max_{v\in\mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} + \delta_m \sum_{k=1}^{K} |\lambda_k| \right].$$

I then propose to choose Λ_F to be the set of points in the neighborhoods of *every* $\tilde{\lambda}_L(\delta_m)$ and $\tilde{\lambda}_U(\delta_m)$. I review the performance of this approach via simulation in Section 7.

When $\delta^* = 0$, i.e., when the plug-in bound is not empty, a researcher may choose $M = 1$ with $\delta_1 = 0$, in which case the procedure reduces to the procedure in Section 5.2. This means that the inference procedure with relaxed bounds generalizes the procedure discussed in Section 5.2.

6 Computation

This section discusses computation of the objects involved in estimation and inference. In particular, it focuses on computation of the two optimization problems in the plug-in lower bound in (13), which apply similarly to other objects, such as the plug-in upper bound in (14), the moment inequalities in (18) and the relaxed plug-in bounds in (22).

6.1 The inner problem

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

$$G_L(\lambda, w) = \min_{v \in \mathcal{V}} \left\{ m(w, v) + \sum_{k=1}^{K} \lambda_k \phi_k(w, v) \right\} \tag{23}$$

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

Evaluation of G_L involves a global minimization of a possibly non-convex function, with G_L required to be evaluated for each $w = W_i$ and for each step of the outer optimization problem. 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, neither of these scenarios is likely to apply.

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

Example 4. 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

$$(R_{it}'V_i)(Y_{it} - R_{it}'V_i), \quad t = 1, \ldots, T, \tag{24}$$

and the entries of the vectors

$$(Z_i', X_i^{t\prime})'(Y_i - R_{it}'V_i), \quad t = 1, \ldots, T, \tag{25}$$

which are, at most, second-order polynomials of V_i.

Example 5. 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

$$(R_{it}'V_i)^3(Y_{it} - R_{it}'V_i), \quad t = 1, \ldots, T, \tag{26}$$

which are fourth-order polynomials of V_i. A researcher may also consider the moment functions in Assumption 3, in which case the additional ϕ_ks are set to be (24) and (25).

In Examples 4 and 5, the moment functions are chosen to 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 that of the parameter of interest. In Examples 4 and 5, I choose (24) to obtain a second-order polynomial and (26) to obtain a fourth-order polynomial as inner objectives. The inner objective polynomial then has a leading coefficient that is positive or negative depending on the signs of λ, which delivers finite inner minimum and maximum that yield finite lower and upper bounds.

The polynomial case can be extended to allow m to be an indicator function of V_i. An indicator function partitions $\mathcal{V}$ into two exclusive sets, with the indicator function constant within each set. A researcher can then compute the global optimum in each partition, followed by the optimum of both. This extension is useful for computing bounds for CDFs of random coefficients, as described in the following example.

Example 6. 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

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

which is an indicator function of V_i. Consider the same set of moment restrictions as in Example 4, 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 $\mathcal{V}_2 = \{(v_1, \ldots, v_{q+p}) \mid v_1 > v\}$, with $m = 1$ on $\mathcal{V}_1$ and $m = 0$ on $\mathcal{V}_2$. The inner objective function is then a second-order polynomial within both $\mathcal{V}_1$ and $\mathcal{V}_2$, for which I can compute the minimum. Finally, I can evaluate the inner objective by taking the smaller optimum of the two in $\mathcal{V}_1$ and $\mathcal{V}_2$.

The next two subsections discuss a fast and precise 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 (23) is expressed in this standard form, (A, b, c) are functions of the data w.

Quadratic polynomials can be solved efficiently using quadratic optimization software. In practice, a researcher can use a heuristic but faster (that is, closed form) method to increase the speed. If A is positive definite, Q attains the global minimum at the solution to the first-order condition, which is given by:

$$\min_{v \in \mathcal{V}} Q(v) = c - \frac{1}{4} b' A^{-1} b. \tag{27}$$

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, all with the same value of Q. Otherwise, the minimum of Q is negative infinity.

In the context of (23), if the data w follows 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, simply using (27) to express (23) in a closed form if and only if A is positive definite; otherwise, the solution is negative infinity.

The heuristic method discussed above applies when $\mathcal{V}$ in (23) is unbounded. In some cases, a researcher may consider restricting $\mathcal{V}$ to be a bounded set, such as by restricting the autoregressive parameter of the AR(1) model to be within $[0, 1]$. In that case, the heuristic method can be modified to incorporate the constraint, using the Lagrange multiplier method. Alternatively, quadratic optimization software can be used.

6.1.2 Global optimization of generic polynomials

When m and ϕ_ks are polynomials of generic order, a closed form solution is not available, but the problem can be solved numerically. The key step is to transform the problem into a convex optimization problem (Lasserre, 2010, 2015). The resulting algorithm is not only fast, but also computes an exact solution. This subsection summarizes the main concept 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 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 $\mathcal{V}$ is equal to the solution of the minimization problem:

$$\min_{P_V \in \mathcal{M}_V,\ \int dP_V = 1} \int \pi(v) dP_V \tag{28}$$

where P_V is a probability distribution on $\mathcal{V}$. (28) is minimized at the point-mass distribution concentrated at the minimizer of $\pi(v)$. Since $\pi(v) = \sum_{j=1}^{J} a_j p_j(v)$, I can rewrite (28) 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:

$$\min_{M_1,\ldots,M_J \in \mathbb{R},\ M_1 = 1} \sum_{j=1}^{J} a_j M_j \qquad \text{subject to} \qquad M_j = \int p_j(v) dP_V \text{ for some } P_V \in \mathcal{M}_V. \tag{29}$$

Except for the fact that the constraint is complicated, (29) is a minimization over $\mathbb{R}^J$ and the objective is linear (thus convex) in the choice variables.

The aim then is to replace the constraint in (29) with a convex constraint that involves only $(M_1, \ldots, M_J)$. The constraint in (29) indicates that $(M_1, \ldots, M_J)$ must represent moments of some underlying distribution, whose necessary condition can be characterized using a matrix[7]. For example, a random variable X must satisfy $\mathbb{E}(X^2) - \mathbb{E}(X)^2 \geq 0$, because $\mathrm{Var}(X)$ must be nonnegative. This is equivalent to the 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 a linear operator $\mathcal{L}$ that maps a polynomial to $\mathbb{R}$ by the 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 follow that:

$$\mathcal{L}(u(v)u(v)') \text{ is positive semidefinite} \tag{30}$$

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

(30) is a convex constraint because the set of positive semidefinite matrices is a convex set. Replacing the constraint in (29) with (30) yields a convex optimization problem:

$$\min_{M_1,\ldots,M_J \in \mathbb{R}} \sum_{j=1}^{J} a_j M_j \qquad \text{subject to} \qquad \mathcal{L}(u(v)u(v)') \text{ is positive semidefinite.} \tag{31}$$

The constraint can be handled more efficiently than a generic convex constraint, meaning that the optimization problem is a semidefinite program (SDP), i.e., 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 (31), the *semidefinite relaxation*, which can be solved quickly and reliably using SDP solvers. The algorithm offers a *certificate of optimality*, a condition for the optimal value of $(M_1, \ldots, M_J)$, satisfying which means that the solution to (31) equals the global optimum. For researchers interested in using the semidefinite relaxation approach to global polynomial optimization, I offer a

[7] See Lasserre (2010, 2015) for the necessary and sufficient condition that involves a sequence of matrices.

general-purpose R package optpoly that implements the approach[8].

The solution to (31) (i.e., the SDP solution) is less than or equal to the solution to (29), since a necessary condition is weaker than the original condition. The semidefinite relaxation approach solves a sequence — or *hierarchy* — of the SDP programs, until the certificate of optimality is obtained, which is known to occur in a finite number of steps under suitable conditions. Even if a researcher does not succeed in solving the hierarchy of the SDPs, the researcher can take an SDP solution as a lower bound for (29), in which case the resulting plug-in bound is a non-sharp but valid bound for θ.

6.2 The outer problem

I now 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 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 2. 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 optimization algorithms such as gradient descent methods.

Concavity of the outer problem comes from the concavity of G_L, and precisely solving the inner problem via the polynomial optimization algorithm is crucial for computational tractability of the outer problem. This differs significantly from Schennach (2014) and Li (2018), who studied generic moment equality models. I focus on random coefficient models and exploit their linear structure to achieve computational tractability for models with large dimensions. If a researcher uses general-purpose global optimization methods that involve random approximation errors (e.g. simulated annealing) to solve the inner problem, then G_L is no longer concave, meaning that fast convex optimization algorithms cannot be used for the outer problem. This is problematic when K is large, which is often the case in applications of random coefficient models.

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

7 Simulation

This section examines the performance of the inference procedure discussed in Section 5. The simulation uses 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}:

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

By 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 determined 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, 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 necessary because the identified set of the parametric model is infeasible to compute, while the identified set for the finite population can be precisely computed using (13) and (14).

I generate a finite population of size 100,000 for each combination of $T \in \{5, 10, 15\}$ and $L \in \{3, 5, 7\}$. T is the length of panel data, and L is the maximum lag of Y_{it} used for the moment conditions. Given L, I use the following set of moment conditions:

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

I also restrict $(\gamma_i, \beta_i) \in \mathcal{V} = [-3, 3] \times [0, 1]$, which is true for the finite populations used in the simulation. I then compute the population identified set for each (T, L).

Then, for each (T, L), I create Monte Carlo replications by sampling $N = 750$ or 1000 observations with replacement from the finite population. I then compute the confidence interval for $\mathbb{E}(\beta_i)$ in each Monte Carlo replication, using the procedure with relaxed bounds in Section 5.3. The grid of $\{\delta_m\}$ is set to be $\delta_m \in \{1.25\delta^*, 1.50\delta^*, 1.75\delta^*, \ldots, 2.75\delta^*, 3\delta^*\}$, which is an equispaced grid of 8 values. For each $\tilde{\lambda}_L(\delta_m)$ and $\tilde{\lambda}_U(\delta_m)$, I sample $P = 25, 50$ or 75 points from their respective neighborhoods, 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 is $8P$. The critical value is computed with 2000 multiplier bootstrap replications.

Tables 1 and 2 present coverage probabilities of the confidence interval for $\mathbb{E}(\beta_i)$ 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 are 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 a higher risk in earnings dynamics accumulate more precautionary savings in order to smooth consumption. Households save more when they experience a positive earnings shock, with the savings used to maintain consumption during a negative earnings shock. Specifying the 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

[9] As Guvenen (2007) points out, this is a stylized version of what is used in the literature, but it still captures features important for the discussion.

earnings dynamics. Consider two earnings processes:

$$\begin{aligned} Y_{it} &= \alpha_i + z_{it} + \varepsilon_{it}, \qquad & z_{it} &= \rho z_{i,t-1} + \eta_{it}, \qquad & &\text{(RIP)} \\ Y_{it} &= \alpha_i + \beta_i h_{it} + z_{it} + \varepsilon_{it}, \qquad & z_{it} &= \rho z_{i,t-1} + \eta_{it}, \qquad & &\text{(HIP)} \end{aligned} \tag{32}$$

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

While there is vast literature on unobserved heterogeneity in β_i and its influence on ρ, there is relatively little work investigating unobserved heterogeneity in ρ itself. Recent studies include Browning, Ejrnaes, and Alvarez (2010) and Alan, Browning, and Ejrnæs (2018), in which unobserved heterogeneity in ρ is given by a factor structure. In this section, I investigate unobserved heterogeneity in ρ by estimating a generalization of (32) where $\rho = \rho_i$. I treat the generalized model as a random coefficient model, meaning that distribution of ρ_i and its dependence on $(\alpha_i, \beta_i, Y_{i0})$ are unrestricted. Distributions of the 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, RIP and HIP have similar estimates of $\mathbb{E}(\rho_i)$ that are significantly less than 1. Confidence intervals for $\mathbb{E}(\rho_i)$ in the two processes have substantial overlap, with both having upper confidence limits of around 0.6 at 90% confidence level. Confidence intervals for the CDF of ρ_i are also similar in 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 substantial heterogeneity in ρ_i, obtaining a lower confidence limit of 0.067 for $\text{Var}(\rho_i)$ in the RIP process at 90% confidence level. This implies a lower confidence limit of 0.258 for the standard deviation of ρ_i.

8.2 Data and model

I use data on U.S. households from the Panel Study of Income Dynamics (PSID) dataset. The data are based on the dataset of Guvenen (2009), who estimated RIP and HIP processes using PSID earnings data. I constructed data of $N = 800$ individuals and $T = 15$ periods from his dataset, details of which can be found in the Online Appendix B.4.

I follow Guvenen (2009) and use potential experience as a measure of experience, $h = \text{age} - \max\{\text{years of schooling}, 12\} - 6$. In addition, since my method requires there to be no multicollinearity in each individual time series (Assumption 2), I remove 40 individuals (that is, 5% of data) with the smallest variations in their reported incomes, yielding a dataset of $N = 760$ and $T = 15$. Estimation results with this removal do not qualitatively differ from results without the removal, which can be found in the Online Appendix B.4.

To apply my method to income processes, I transform two models in (32) to random coefficient models. I first remove ε_{it} from Y_{it} using a simulation-based de-noising method inspired by Arellano and Bonhomme (2021), described in the Online Appendix B.5, obtaining pseudo-observations of the permanent incomes, $\tilde{Y}_{it}$. I then write, using (32):

$$\begin{aligned} \tilde{Y}_{it} &= \alpha_i + z_{it}, & z_{it} &= \rho z_{i,t-1} + \eta_{it}, & &\text{(RIP)} \\ \tilde{Y}_{it} &= \alpha_i + \beta_i h_{it} + z_{it}, & z_{it} &= \rho z_{i,t-1} + \eta_{it}. & &\text{(HIP)} \end{aligned} \tag{33}$$

Estimation results with this de-noising do not qualitatively differ from results without de-noising, which can be found in the Online Appendix B.5. I then quasi-difference (33) to transform them to random coefficient models. Quasi-differencing each of (33) yields, respectively:

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

These are standard random coefficient models. I assume that h_{it} is a strictly exogenous regressor, meaning that years of schooling is strictly exogenous.

8.3 Strategy for estimation and inference

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

Example 4. In particular, I use for the RIP process:

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

for $s = 0, \ldots, 5$. I use for the HIP process:

$$\mathbb{E}((\tilde{\alpha}_i + \tilde{\beta}_i h_{it} + \rho_i \tilde{Y}_{i,t-1})\eta_{it}) = 0, \quad \mathbb{E}(\eta_{it}) = 0, \quad \mathbb{E}(\tilde{Y}_{i,t-1-r}\eta_{it}) = 0, \quad \mathbb{E}(h_{i,t-s}\eta_{it}) = 0,$$

for $r = 0, \ldots, 5$ and $s = -5, \ldots, -1, 0, 1, \ldots, 5$. These make the inner objective a second order polynomial. I then solve the inner optimization problem in a closed form.

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

$$\mathbb{E}((\tilde{\alpha}_i + \rho_i \tilde{Y}_{i,t-1})^3 \eta_{it}) = 0, \qquad \mathbb{E}(\tilde{\alpha}_i^{k_\alpha} \rho_i^{k_\rho} \eta_{it}) = 0,$$

for $0 \leq k_\alpha + k_\rho \leq 2$ where k_α and k_ρ are integers. Additional moments for the HIP process are:

$$\mathbb{E}((\tilde{\alpha}_i + \tilde{\beta}_i h_{it} + \rho_i \tilde{Y}_{i,t-1})^3 \eta_{it}) = 0, \qquad \mathbb{E}(\tilde{\alpha}_i^{k_\alpha} \tilde{\beta}_i^{k_\beta} \rho_i^{k_\rho} \eta_{it}) = 0,$$

for $0 \leq k_\alpha + k_\beta + k_\rho \leq 2$ where k_α, k_β and k_ρ are integers. The first additional moment restriction in both models was stated in Example 5, which makes the inner objective a fourth-order polynomial. The second additional moment restriction then adds lower-order terms to the inner objective. These additional restrictions yield finite lower and upper bounds on the second moments of $(\tilde{\alpha}_i, \tilde{\beta}_i, \rho_i)$. I then solve the inner problem using the SDP method with hierarchy of length two.

With these moment restrictions, I compute confidence intervals using the procedure in Section 5.3. Tuning parameters for the inference procedure are the same as those in the simulations, sampling $P = 50$ points in the neighborhood of each $\tilde{\lambda}_L(\delta_m)$ and $\tilde{\lambda}_U(\delta_m)$.

8.4 Estimation results

Confidence intervals for $\mathbb{E}(\rho_i)$ and $\text{Var}(\rho_i)$ are given in Table 3. Both models estimate $\mathbb{E}(\rho_i)$ to be significantly less than 1, in contrast to the models with homogeneous ρ where the literature estimates $\rho \approx 1$ for the RIP process. Moreover, the confidence intervals of the two models demonstrate substantial overlap, having similar upper confidence limits. This suggests that specifying homogeneous or heterogeneous β does not lead to serious misspecification when ρ is allowed to be heterogeneous.

The confidence interval for $\text{Var}(\rho_i)$ suggests substantial heterogeneity in ρ_i, with a lower confidence limit of 0.067 for the RIP process implying a standard deviation of 0.258. Similar evidence is observed from confidence intervals for the CDF of ρ_i in Table 4. Confidence intervals for the CDF of ρ_i in the RIP process suggest that at least 15% of households have $\rho_i \leq 0.5$, while another 15%, at least, of households have $\rho_i > 0.5$. Confidence intervals for the two CDFs show substantial overlap. These suggest substantial unobserved heterogeneity in the earnings risk that households face, which is a key source of heterogeneity in consumption and savings behaviors. These highlight the importance of allowing for heterogeneity in ρ_i in modeling income processes that reflect features of real data.

9 Conclusion

This paper studies identification and estimation of dynamic random coefficient models in a short panel context. The model extends the widely used panel data linear model with fixed effects (Arellano and Bond, 1991; Blundell and Bond, 1998), by allowing for individual-specific coefficients and intercept. I show that the model is not point-identified but rather partially identified, and I characterize sharp identified sets of the parameters of interest using the dual representation of the infinite-dimensional linear program. I propose a computationally feasible estimation procedure whose computational feasibility is achieved using a fast and precise algorithm for global polynomial optimization, which also yields a computationally feasible inference procedure based on testing many moment inequalities.

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 earnings process. Estimates for variance and CDF of earnings persistence suggest a substantial degree of unobserved heterogeneity, which is a key source of heterogeneity in consumption and savings behaviors.

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 their guidance and support. I thank Manuel Arellano, Timothy Armstrong, Antonio Galvao, Jack Light, Azeem Shaikh, and Panagiotis Toulis for helpful discussions and comments. I also thank participants of the 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.

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 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.

———. 2021. "Recovering latent variables by matching." *Journal of the American Statistical Association* :1–14.

Bai, Yuehao, Andres Santos, and Azeem M Shaikh. 2022. "A two-step method for testing many moment inequalities." *Journal of Business & Economic Statistics* 40 (3):1070–1080.

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* .

———. 2022. "Feedback in panel data models." *Journal of Econometrics* 226 (1):4–20.

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.

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

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

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

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* .

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.

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

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

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.

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.

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." *Econometrica* 87 (5):1475–1505.

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

Proof of Proposition 1. This proof is an application of the general result in the Online Appendix B.2. Assume the regularity conditions stated as Assumption 7 in the Online Appendix B.2, where the item (iv) follows from the assumption that the joint density of $(Y_{i0}, Y_{i1}, Y_{i2}, \gamma_i, \beta_i)$ is strictly positive (Anderson, 1983, Theorem 9). Also, 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 a contradiction. Lemma 3 in the Online Appendix B.2 tells that, if $\mathbb{E}(\beta_i)$ is point-identified, it follows that:

$$f^*(Y_{i0}, Y_{i1}, Y_{i2}) + g_1^*(\gamma_i, \beta_i, Y_{i0})\varepsilon_{i1} + g_2^*(\gamma_i, \beta_i, Y_{i0}, Y_{i1})\varepsilon_{i2} = \beta_i \tag{34}$$

almost surely in $(\gamma_i, \beta_i, Y_{i0}, Y_{i1}, Y_{i2})$, where $f^* : \mathcal{C}_0^3 \mapsto \mathbb{R}$, $g_1^* : \mathcal{C}_0^3 \mapsto \mathbb{R}$ and $g_2^* : \mathcal{C}_0^4 \mapsto \mathbb{R}$ are linear functionals on the spaces of finite and countably additive signed Borel measures that are absolutely continuous with respect to the Lebesgue measure. Substituting $\varepsilon_{it} = Y_{it} - \gamma_i - \beta_i Y_{i,t-1}$ in (34) yields, almost surely in $(\gamma_i, \beta_i, Y_{i0}, Y_{i1}, Y_{i2})$:

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

Consider any $\gamma, \tilde{\gamma}, \beta, y_0, y_1, y_2 \in \mathcal{C}_0$ such that $\gamma \neq \tilde{\gamma}$. I evaluate (35) at $(\gamma_i, \beta_i, Y_{i0}, Y_{i1}, Y_{i2}) = (\gamma, \beta, y_0, y_1, y_2)$ and at $(\tilde{\gamma}, \beta, y_0, y_1, y_2)$, and I take the difference, which yields:

$$\begin{aligned} &(y_1 - \tilde{\gamma} - \beta y_0)\triangle_{\tilde{\gamma},\gamma} g_1^* - (\tilde{\gamma} - \gamma) g_1^*(\gamma, \beta, y_0) \\ &+ (y_2 - \tilde{\gamma} - \beta y_1)\triangle_{\tilde{\gamma},\gamma} g_2^* - (\tilde{\gamma} - \gamma) g_2^*(\gamma, \beta, y_0, y_1) = 0 \end{aligned} \tag{36}$$

where $\triangle_{\tilde{\gamma},\gamma} g_1^* \equiv g_1^*(\tilde{\gamma}, \beta, y_0) - g_1^*(\gamma, \beta, y_0)$ and $\triangle_{\tilde{\gamma},\gamma} g_2^* \equiv g_2^*(\tilde{\gamma}, \beta, y_0, y_1) - g_2^*(\gamma, \beta, y_0, y_1)$.

In (36), y_2 appears only in the third term. Also, (36) must hold almost surely for all $\gamma, \tilde{\gamma}, \beta, y_0, y_1, y_2 \in \mathcal{C}_0$ such that $\gamma \neq \tilde{\gamma}$, and in particular for any $y_2 \in \mathcal{C}_0$. This implies that, almost surely:

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

which means that g_2^* is almost surely a constant function over γ:

$$g_2^*(\gamma, \beta, y_0, y_1) = g_2^*(\beta, y_0, y_1). \tag{37}$$

If not, i.e., if $\triangle_{\tilde{\gamma},\gamma} g_2^* \neq 0$ on a subset of $\mathcal{C}_0^6$ with positive Lebesgue measure, one can change the value of y_2 without changing $(\gamma, \tilde{\gamma}, \beta, y_0, y_1)$ within this subset to violate (36) with a positive measure.

Next, consider any $\gamma, \beta, \tilde{\beta}, y_0, y_1, y_2 \in \mathcal{C}_0$ such that $\beta \neq \tilde{\beta}$. I evaluate (35) at $(\gamma, \beta, y_0, y_1, y_2)$ and $(\gamma, \tilde{\beta}, y_0, y_1, y_2)$, and I take the difference:

$$\begin{aligned}&(y_1 - \gamma - \tilde{\beta} y_0) \triangle_{\tilde{\beta},\beta} g_1^* - (\tilde{\beta} - \beta) y_0 g_1^*(\gamma, \beta, y_0)\\ &+ (y_2 - \gamma - \tilde{\beta} y_1) \triangle_{\tilde{\beta},\beta} g_2^* - (\tilde{\beta} - \beta) y_1 g_2^*(\beta, y_0, y_1) = \tilde{\beta} - \beta\end{aligned} \tag{38}$$

where $\triangle_{\tilde{\beta},\beta} g_1^* \equiv g_1^*(\gamma, \tilde{\beta}, y_0) - g_1^*(\gamma, \beta, y_0)$ and $\triangle_{\tilde{\beta},\beta} g_2^* = g_2^*(\tilde{\beta}, y_0, y_1) - g_2^*(\beta, y_0, y_1)$. In (38), y_2 appears only in the third term. This implies $g_2^*(\beta, y_0, y_1) = g_2^*(y_0, y_1)$ almost surely, similarly to the argument for (37). Then (36) simplifies to:

$$(y_1 - \tilde{\gamma} - \beta y_0) \triangle_{\tilde{\gamma},\gamma} g_1^* - (\tilde{\gamma} - \gamma) g_1^*(\gamma, \beta, y_0) - (\tilde{\gamma} - \gamma) g_2^*(y_0, y_1) = 0. \tag{39}$$

Let $\gamma, \tilde{\gamma}, \hat{\gamma} \in \mathcal{C}_0$ be such that $\hat{\gamma} - \tilde{\gamma} = \tilde{\gamma} - \gamma$. I evaluate (39) at $(\gamma, \tilde{\gamma}, \beta, y_0, y_1)$ and $(\tilde{\gamma}, \hat{\gamma}, \beta, y_0, y_1)$, and I take the difference:

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

In (40), y_1 appears only in the first term, which implies $\triangle_{\hat{\gamma},\tilde{\gamma}} g_1^* - \triangle_{\tilde{\gamma},\gamma} g_1^* = 0$ almost surely, similarly to the argument for (37). Then (40) simplifies to:

$$(\hat{\gamma} - \tilde{\gamma}) \triangle_{\tilde{\gamma},\gamma} g_1^* + (\tilde{\gamma} - \gamma) \triangle_{\tilde{\gamma},\gamma} g_1^* = 0, \tag{41}$$

which implies $\triangle_{\tilde{\gamma},\gamma} g_1^* = 0$ since $\hat{\gamma} - \tilde{\gamma} = \tilde{\gamma} - \gamma \neq 0$. This implies that g_1^* is almost surely a constant function over γ, i.e., $g_1^*(\gamma, \beta, y_0) = g_1^*(\beta, y_0)$. Then (38) simplifies to:

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

Let $\beta, \tilde{\beta}, \hat{\beta} \in \mathcal{C}_0$ be such that $\hat{\beta} - \tilde{\beta} = \tilde{\beta} - \beta$. Evaluating (42) at $(\gamma, \hat{\beta}, \tilde{\beta}, y_0, y_1, y_2)$ and at $(\gamma, \tilde{\beta}, \beta, y_0, y_1, y_2)$ and taking the difference yields $g_1^*(\beta, y_0) = g_1^*(y_0)$, in a similar way to the argument for $g_1^*(\gamma, \beta, y_0) = g_1^*(\beta, y_0)$ from (40). Then (35) 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 must coincide on both sides. This means that $-g_1^* - g_2^* = 0$ (from the coefficients on γ) and $-y_0 g_1^* - y_1 g_2^* = 1$ (from the coefficients on β). Solving these 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 a contradiction. □

Proof of Lemma 1. As discussed in the proof of Proposition 1, Lemma 3 in the Online Appendix B.2 tells us that if $\mathbb{E}(\beta_i)$ is point-identified, there exists (f^*, g_1^*, g_2^*) such that (34) holds almost surely on $\mathcal{C}_0^5$. Then (34) implies $S^*(Y_{i0}, Y_{i1}, Y_{i2}) = f^*(Y_{i0}, Y_{i1}, Y_{i2})$ because:

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

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. □

Proof of Theorem 1. It suffices to show that $[L, U]$ is the sharp bound when Assumption 3 is replaced by (9). In what follows, I show that U is the sharp upper bound under (9). The same argument applies to L. This proof is an application of Theorem 2 in Section 4.

According to Theorem 2, the sharp upper bound is given by:

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

where μ has the same dimension as R_{it}, and λ is scalar. With the notation of $\mathcal{R}_i$ and $\mathcal{Y}_i$ in the statement of Theorem 1, I can write the above concisely as

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

The inner maximization problem optimizes a quadratic polynomial in v. This quadratic maximization problem has a closed form solution if $\lambda > 0$, and it diverges to $+\infty$ almost surely if $\lambda \leq 0$, as discussed in Section 6.1.1. For $\lambda > 0$, the closed form solution yields:

$$\min_{\lambda>0,\, \mu} \mathbb{E}\left(\mu'\mathcal{Y}_i + \frac{1}{4\lambda}\left[e + \lambda\mathcal{Y}_i - \mathcal{R}_i\mu\right]' \mathcal{R}_i^{-1}\left[e + \lambda\mathcal{Y}_i - \mathcal{R}_i\mu\right]\right). \tag{43}$$

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

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

The optimal μ that solves this first-order condition is $\mu^* = \mathbb{E}(\mathcal{R}_i)^{-1}[e - \lambda\mathbb{E}(\mathcal{Y}_i)]$. I substitute this into (43) and solve (43) 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 that solves this first-order condition is:

$$\lambda^* = \sqrt{\frac{e'\mathbb{E}(\mathcal{R}_i^{-1})e - e'\mathbb{E}(\mathcal{R}_i)^{-1}e}{\mathbb{E}(\mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i) - \mathbb{E}(\mathcal{Y}_i)'\mathbb{E}(\mathcal{R}_i)^{-1}\mathbb{E}(\mathcal{Y}_i)}}. \tag{44}$$

Substituting (44) into (43) yields the expression of $\tilde{U}$ in Theorem 1.

The numerator and denominator in (44) 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 show this, define the functions $E(\mathcal{R}_i) = e'\mathcal{R}_i^{-1}e$ and $D(\mathcal{Y}_i, \mathcal{R}_i) = \mathcal{Y}_i'\mathcal{R}_i^{-1}\mathcal{Y}_i$, and apply the following proposition to E and D. □

Proposition 4 (Kiefer, 1959, Lemma 3.2)**.** *For an integer $l > 0$, let $A_1, \ldots, A_l$ be $n \times m$ matrices and $B_1, \ldots, B_l$ be nonsingular positive definite and symmetric $n \times n$ matrices. Let $a_1, \ldots, a_l$ be positive real numbers such that $\sum_k a_k = 1$. Then*

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

where '$\geq$' is the partial ordering defined in terms of positive semidefinite matrices. In addition, the equality holds if and only if $B_1^{-1}A_1 = \ldots = B_l^{-1}A_l$.

Proof of Theorem 2. In what follows, I prove that (11) is the dual representation of (10). The proof is a direct application of the duality theorem of linear programming for topological vector spaces (Anderson, 1983). The same argument applies to (12).

To apply the theorem, I first rewrite (10) as a standard form of 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*

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

Let $\mathcal{M}_W$ be the 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 $g=(g_1,\ldots,g_K,P_g)$ and $h=(\lambda_1,\ldots,\lambda_K,f_h)$ be their generic elements. Note that $\mathcal{H}$ is the dual space of $\mathcal{G}$. Define the dual pairing

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

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 (thus 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 that:

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

Then:

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

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

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

(45) shows that A^* is the adjoint of A. With these notations, I rewrite (10) as a standard form of linear programming:

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

where $c = (0, \ldots, 0, P_W)$. I then apply the strong duality theorem (Anderson, 1983, Theorem 6) under Assumption 5 and the continuity of A, which tells that the optimal solution to (46) is equal to the solution to:

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

which I can write more concretely as:

$$\max_{\lambda_1, \ldots, \lambda_K \in \mathbb{R},\, f_h \in \overline{\mathcal{F}}_W} \int f_h dP_W \qquad \text{subject to} \qquad \sum_{k=1}^{K} \lambda_k \phi_k + f_h \leq m. \tag{47}$$

Now I show that (47) simplifies to (11). I rearrange the constraint of (47):

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

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 (47) maximizes the expectation of f_h, the optimal f_h^* for a fixed $(\lambda_1, \ldots, \lambda_K)$ is given by:

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

almost surely in P_W. If not, i.e., if $f_h^*(w)$ is strictly less than the right-hand side of (48) with positive probability in P_W, one can increase the value of the objective by increasing f_h^* on a set of positive probability. Next, I substitute (48) into (47), which yields:

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

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

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

which is the expression in (11). □

Proof of Lemma 2 As in (47) in the proof of Theorem 2, the sharp lower bound of θ is given by

$$\max_{\lambda_1,\ldots,\lambda_K \in \mathbb{R},\, f_h \in \overline{\mathcal{F}}_W} \int f_h dP_W \qquad \text{subject to} \qquad \sum_{k=1}^{K} \lambda_k \phi_k + f_h \leq m \tag{49}$$

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

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

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

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

Then the solution $(\lambda_1^l, \ldots, \lambda_K^l, S^l)$ to (49) satisfies its constraint $\sum_{k=1}^{K} \lambda_k^l \phi_k + S^l \leq m$ with strict inequality 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 (50) satisfies its constraint $\sum_{k=1}^{K} \lambda_k^u \phi_k + S^u \geq m$ with strict inequality on a set with positive Lebesgue measure on $\mathcal{W} \times \mathcal{V}$. Then:

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

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 a contradiction since θ is assumed to be point-identified.

Conversely, suppose there exists $(S^*, \lambda_1^*, \ldots, \lambda_K^*)$ such that $\sum_{k=1}^{K} \lambda_k^* \phi_k + S^* = m$. Then:

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

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

Proof of Proposition 2. It suffices to show that G_L is concave in λ. 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 $t \in [0,1]$ and $w \in \mathcal{W}$:

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

This is the definition of concavity. □

Proof of Proposition 3. When L exists and is finite, Proposition 2 implies that $L(\lambda)$ is concave. Then $\hat{L}$ uniformly converges to L on any compact set $K_0 \subseteq \mathbb{R}^K$, as in the proof of Theorem 2.7 in Newey and McFadden (1994):

$$\sup_{\lambda \in K_0} |\hat{L}(\lambda) - L(\lambda)| \xrightarrow{p} 0. \tag{51}$$

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

$$\begin{aligned}
|L(\lambda_0) - \hat{L}(\hat{\lambda})| &\leq L(\lambda_0) - L(\hat{\lambda}) + |L(\hat{\lambda}) - \hat{L}(\hat{\lambda})| && \text{(triangle inequality)} \\
&= \hat{L}(\lambda_0) - L(\hat{\lambda}) + |L(\hat{\lambda}) - \hat{L}(\hat{\lambda})| + o_p(1) && \text{(by (51))} \\
&\leq \hat{L}(\hat{\lambda}) - L(\hat{\lambda}) + |L(\hat{\lambda}) - \hat{L}(\hat{\lambda})| + o_p(1) && \text{($\hat{\lambda}$ is argmax)} \\
&\leq 2|L(\hat{\lambda}) - \hat{L}(\hat{\lambda})| + o_p(1) = o_p(1). && \text{(by (51))}
\end{aligned}$$

Let Λ_C be a compact set containing $\Lambda_0 \equiv \{\lambda \in \mathbb{R}^K \mid \lambda = \text{argmax}_\lambda L(\lambda)\}$ such that its boundary Λ_C^B satisfies $\max_{\lambda \in \Lambda_C^B} L(\lambda) < L(\lambda_0) - \varepsilon$ for some $\varepsilon > 0$. Let $\tilde{\lambda}$ be the maximizer of $\hat{L}$ on Λ_C, i.e., $\tilde{\lambda} \equiv \text{argmax}_{\lambda \in \Lambda_C} \hat{L}(\lambda)$. Then, by (51), $|\hat{L}(\lambda) - L(\lambda)| < \varepsilon/3$ uniformly on Λ_C with probability approaching to one, so that $\hat{L}(\tilde{\lambda}) > \max_{\lambda \in \Lambda_C^B} \hat{L}(\lambda)$.

Now, for any λ' outside of Λ_C, there exists $t \in [0,1]$ such that $t\tilde{\lambda} + (1-t)\lambda' \in \Lambda_C^B$, so that $\hat{L}(\tilde{\lambda}) > \hat{L}(t\tilde{\lambda} + (1-t)\lambda')$. Also, the concavity of $\hat{L}$ implies that $\hat{L}(t\tilde{\lambda} + (1-t)\lambda') \geq t\hat{L}(\tilde{\lambda}) + (1-t)\hat{L}(\lambda')$. Combining the two yields $(1-t)\hat{L}(\tilde{\lambda}) \geq (1-t)\hat{L}(\lambda')$, implying that $\tilde{\lambda}$ is the maximizer of $\hat{L}$ not only on Λ_C but also on $\mathbb{R}^K$, meaning that $\tilde{\lambda} = \hat{\lambda}$. Therefore, $\hat{\lambda}$ is contained in the compact set Λ_C with probability approaching to one. □

$(T = 5)$	$L = 3$	$L = 5$	$L = 7$
$P = 25$	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 1: Coverage probabilities of the inference procedures with the sample size of $N = 750$. The 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 2: Coverage probabilities of the inference procedures with the sample size of $N = 1000$. The nominal coverage probability is 0.9.

	$\mathbb{E}(\rho_i)$	$\text{Var}(\rho_i)$
RIP process	[0.456, 0.615]	[0.067, 0.292]
HIP process	[0.264, 0.583]	[0.000, 0.701]

Table 3: Confidence intervals for $\mathbb{E}(\rho_i)$ and $\text{Var}(\rho_i)$ of the RIP and the HIP processes. The nominal coverage probability is 0.9.

$\mathbb{P}(\rho_i \leq r)$	RIP process	HIP process
$r = 0.0$	[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 4: Confidence intervals for $\mathbb{P}(\rho_i \leq r)$ of the RIP and the HIP processes. The nominal coverage probability is 0.9.

THIS PAGE IS INTENTIONALLY LEFT BLANK

B Online Appendix

B.1 Extension to multivariate random coefficient models

Results from this paper extend to a multivariate version of (1), a system of random coefficient models:

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

where $\mathbf{Y}_{it}$ is a $D \times 1$ vector of dependent 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). The following is an example of a multivariate random coefficient model.

Example 7 (Joint model of household earnings and consumption behavior)**.** A researcher can combine (3) and (4) and consider a joint lifecycle model of earnings and consumption behavior. When the time t consumption equation and the time $t+1$ earnings equation are combined, a system of random coefficient models is obtained:

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

which can be written in matrix form:

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

In this model, γs and βs can freely correlate among themselves and with (Y_{i0}, A_{i1}), allowing for correlation between earnings and consumption processes.

B.2 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:

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

where $\phi_k, \psi_k : \mathcal{W} \times \mathcal{V} \mapsto \mathbb{R}$ are moment functions, $A_{i1}, \ldots, A_{iK_C}$ are subvectors of (W_i, V_i) and $K_U, K_C \in \mathbb{N}$ 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}, so that (A_{ik}, A'_{ik}) is a rearrangement of (W_i, V_i). I then 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 $\mathcal{A}_k \times \mathcal{A}'_k$ is the rearrangement of $\mathcal{W} \times \mathcal{V}$ according to (A_{ik}, A'_{ik}).

I assume the following regularity conditions:

Assumption 7. The following conditions hold.

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

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

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

(iv) The following set is closed:

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

Assumption 7 (ii) is restrictive, but it is useful enough for proving Proposition 1. The rest of the conditions are similar to 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 density, is strictly positive on $\mathcal{W} \times \mathcal{V}$ (Anderson, 1983, Theorem 9).

Under these assumptions, I obtain the following theorem and the lemma, which are counterparts of Theorem 2 and Lemma 2, respectively, characterizing the identified set I of θ and providing a necessary and sufficient condition for point-identification of θ.

Theorem 3. *Suppose that 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*

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

and

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

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

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

Let $\mathcal{M}_{W \times V}$ 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 the observed data distribution P_W. The identified set I is then defined by

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

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, while the third line represents conditional moment restrictions.

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

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

Now I show that (52) is the dual representation of (54), by a direct application of the duality theorem of linear programming for topological vector spaces (Anderson, 1983). 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 $\mathcal{W}$. I 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 $\mathcal{G} = \mathcal{H} = \mathbb{R}^K \times L^2(\mathcal{A}_1) \times \ldots \times L^2(\mathcal{A}_{K_C}) \times L^2(\mathcal{W})$. I denote their generic elements as $g = (g_1, \ldots, g_{K_U}, \overline{g}_1, \ldots, \overline{g}_{K_C}, f_g)$ 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 a linear map $A : \mathcal{M}_{W \times V} \mapsto \mathcal{G}$ by

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

A is a bounded (thus 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:

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

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

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

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

Then, as in the proof of Theorem 2, I apply the strong duality theorem (Anderson, 1983, Theorem 6) under Assumption 7 and the continuity of A, which tells that the optimal solution to (54) is equal to the solution to:

$$\max_{\lambda_1,\ldots,\lambda_{K_U},\mu_1,\ldots,\mu_{K_C},f_h} \int f_h(w)p_w(w)dw \qquad \text{subject to} \qquad \sum_{k=1}^{K_U}\lambda_k\phi_k + \sum_{k=1}^{K_C}\mu_k\psi_k + f_h \leq m. \tag{56}$$

Simplifying (56), as in the proof of Theorem 2, yields the expression in (52).

□

Lemma 3. *Suppose that the 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* $\mathcal{M}_W$ *(which is 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 (56) in 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 follows the proof of Theorem 3. Similarly, the sharp upper bound of θ is given by

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

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

□

B.3 Estimation and inference under over-identification

In practice, the plug-in bound $[\hat{L}, \hat{U}]$ defined in (13) and (14) may yield an empty set, in which case $\hat{L}$ diverges to $+\infty$ and $\hat{U}$ diverges to $-\infty$. This happens when the empirical data distribution $\hat{P}_W$ does not satisfy the moment restrictions, which may occur even if the population data distribution P_W satisfies the restrictions. In this case, the empirical version of (10) (where P_W is replaced with $\hat{P}_W$) does not have a feasible solution, resulting in an empty plug-in bound. This scenario is comparable with over-identification in the generalized method of moments (GMM) estimation, where the GMM objective may be strictly positive in the sample even if the moments are correctly specified.

There are two approaches for addressing this issue. First, a researcher may obtain a point estimate that minimizes the distance between the model and the data. Second, the researcher may directly obtain a confidence interval without insisting on a point estimate, assuming that the model is correctly specified.

For the first approach, consider the following relaxation of the moment restrictions:

$$|\mathbb{E}(\phi_k(W_i, V_i))| \leq \delta, \quad k = 1, \ldots, K, \tag{57}$$

where $\delta \geq 0$, which reduces to Assumption 4 when $\delta = 0$. This can be considered an absolute-value GMM criterion. The following proposition explains how to compute the smallest δ that allows (20) to hold with the empirical distribution.

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

$$\min_{P \in \mathcal{M}_{W \times V},\, P \geq 0,\, \delta \geq 0} \delta \qquad \textit{subject to} \qquad \begin{aligned} &\left| \int \phi_k(W_i, V_i) dP \right| \leq \delta, \quad k = 1, \ldots, K, \\ &\int P(w, dV_i) = \hat{P}_W(w) \ \textit{ for all } w \in \mathcal{W}, \end{aligned} \tag{58}$$

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

$$\max_{\lambda_1, \ldots, \lambda_K} \frac{1}{N} \sum_{i=1}^{N} \min_{v \in \mathcal{V}} \left\{ \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} \qquad \textit{subject to} \qquad \sum_{k=1}^{K} |\lambda_k| \leq 1. \tag{59}$$

Proof. I can rewrite (58) as:

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

I can then replicate the proof of Theorem 2, obtaining (59) as the simplified dual representation of (58). □

Proposition 5 shows that a researcher can find the smallest δ by solving (59), which is similar to the plug-in bound problem. One difference is that (59) is a constrained optimization problem; however, the constraint has a simple structure whose Jacobian can be derived in closed form.

Let δ^* be the solution to (59), and let $\delta \geq \delta^*$. I then compute the lower bound $\tilde{L}$ under the relaxation (57) by computing the plug-in lower bound with a negative L^1 penalty on λ, with δ being the penalty multiplier:

$$\tilde{L} = \max_{\lambda_1,\ldots,\lambda_K} \left[\frac{1}{N} \sum_{i=1}^{N} \min_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} - \delta \sum_{k=1}^{K} |\lambda_k| \right]. \tag{60}$$

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

Proposition 6. *Given the sample $(W_1, \ldots, W_N)$ and given $\delta \in \mathbb{R}$, consider the linear program that finds the smallest value of $\theta = \mathbb{E}(m(W_i, V_i))$ that satisfies (57):*

$$\min_{P \in \mathcal{M}_{W \times V},\, P \geq 0} \int m(W_i, V_i) dP \qquad \textit{subject to} \qquad \begin{aligned} &\left| \int \phi_k(W_i, V_i) dP \right| \leq \delta, \quad k = 1, \ldots, K, \\ &\int P(w, dv) = \hat{P}_W(w) \ \textit{for all } w \in \mathcal{W}, \end{aligned} \tag{61}$$

where $\hat{P}_W$ is the empirical distribution of W_i constructed from $(W_1, \ldots, W_N)$. Then its solution is equal to $\tilde{L}$ defined in (60).

Proof. I can rewrite (61) as:

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

I can then replicate the proof of Theorem 2, obtaining (60) as the simplified dual representation of (61). □

Proposition 6 shows that (60) equals the smallest value of θ among the distributions whose absolute-value GMM criterion defined in (57) 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 $[\tilde{L}, \tilde{U}]$ becomes a point.

In practice, due to machine precision or the stopping criterion of the optimization algorithm, the numerical solution to (59) might be strictly smaller than its analytical solution. To resolve this problem, a researcher may choose δ to be strictly larger than the numerical solution of δ^*, in which case (60) computes the smallest value of θ among the distributions that attain the *near-minimum* of the absolute-value GMM criterion. If the minimizer distribution is unique, the relaxed plug-in bound with $\delta > \delta^*$ becomes a small interval instead of a point.

Although (60) resolves the empty set problem, it has two drawbacks. First, it is an ad hoc approach, with no formal justification for why the relaxation of moment conditions is a constructive idea. Second, the procedure may yield a point identified set (or a small interval) even if the model is partially identified. The literature deals with the second problem by choosing δ that is substantially larger than δ^* (Mogstad, Santos, and Torgovitsky, 2018), but the question of how much larger it should be remains unresolved. In the remainder of this subsection, I discuss a more principled approach, which is to directly compute a confidence interval without insisting on a point estimate.

Note that the 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 the plug-in bound problems. The inference procedure is valid regardless of whether the plug-in bound is empty; the issue is that no guidance exists 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 to use the relaxed plug-in bounds for choosing Λ_F. The procedure consists of three steps. The first step solves (59) and finds the minimum δ^*. The second step considers a grid of positive real numbers $\{\delta_1, \ldots, \delta_M\}$ such that $\delta_m > \delta^*$ for all $m \in \{1, \ldots, M\}$. Then, for each δ_m, the relaxed plug-in bound is computed:

$$\begin{aligned}\tilde{\lambda}_L(\delta_m) &= \underset{\lambda_1,\ldots,\lambda_K}{\operatorname{argmax}} \left[\frac{1}{N} \sum_{i=1}^{N} \min_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} - \delta_m \sum_{k=1}^{K} |\lambda_k| \right]. \\ \tilde{\lambda}_U(\delta_m) &= \underset{\lambda_1,\ldots,\lambda_K}{\operatorname{argmin}} \left[\frac{1}{N} \sum_{i=1}^{N} \max_{v \in \mathcal{V}} \left\{ m(W_i, v) + \sum_{k=1}^{K} \lambda_k \phi_k(W_i, v) \right\} + \delta_m \sum_{k=1}^{K} |\lambda_k| \right].\end{aligned} \tag{62}$$

The third step then chooses Λ_F to be the set of points in the neighborhoods of *every* $\tilde{\lambda}_L(\delta_m)$ and $\tilde{\lambda}_U(\delta_m)$. In the simulation and the 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 review performance of this approach via simulation in Section 7.

When $\delta^* = 0$, i.e., when the plug-in bound is not empty, a researcher may choose $M = 1$ with $\delta_1 = 0$, in which case the procedure reduces to the procedure in Section 5.2. This means that the inference procedure with relaxed bounds generalizes the procedure discussed in Section 5.2.

B.4 Construction of the dataset

I use data on U.S. households from the Panel Study of Income Dynamics (PSID) dataset. I use the dataset of Guvenen (2009), who estimated RIP and HIP processes using the PSID earnings data of male heads 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:

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

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 remove 40 individuals (that is, 5% of data) with the smallest variations in their reported incomes, giving a dataset of $N = 760$ and $T = 15$. Estimation results with this removal do not qualitatively differ from the results with full data. Tables 5 and 6 present confidence intervals for $\mathbb{E}(\rho_i)$, $\text{Var}(\rho_i)$ and the CDF of ρ_i, computed without removing observations. The confidence intervals do not qualitatively differ from Tables 3 and 4 in the main text, in that the upper confidence limit of $\mathbb{E}(\rho_i)$ is significantly less than 1, the lower confidence limit of $\text{Var}(\rho_i)$ is strictly positive for the RIP process, and the CDF of ρ_i has confidence limits away from 0 and 1.

B.5 Simulation-based de-noising method

Before describing how I removed ε_{it} from Y_{it}, I first describe the simulation-based de-noising method proposed in Arellano and Bonhomme (2021). They considered a model

$$Z = X + \varepsilon \tag{63}$$

where all variables are scalar[10] and X is independent of ε. In this model, Z is observed, but X and ε are not observed. Instead, the distribution of ε is known. The objective of Arellano and Bonhomme (2021) is to obtain pseudo-observations from the distribution of X, given the observations from Z and the knowledge on the distribution of ε.

Let $\mathbb{P}_Z$, $\mathbb{P}_X$ and $\mathbb{P}_\varepsilon$ be the probability distributions of Z, X and ε, respectively. Let $\mathbb{P}_{X+\varepsilon}$ be the distribution of $X+\varepsilon$, which is equal to the convolution of $\mathbb{P}_X$ and $\mathbb{P}_\varepsilon$. The second-order Wasserstein distance between Z and $X+\varepsilon$, denoted by $W_2(\mathbb{P}_Z, \mathbb{P}_{X+\varepsilon})$, is defined by:

$$W_2(\mathbb{P}_Z, \mathbb{P}_{X+\varepsilon}) = \left(\min_{\pi \in \Pi(\mathbb{P}_Z, \mathbb{P}_{X+\varepsilon})} \int ||z - \hat{z}||^2 d\pi(z, \hat{z}) \right)^{1/2}, \tag{64}$$

where $\Pi(\mathbb{P}_Z, \mathbb{P}_{X+\varepsilon})$ is the set of *couplings* of $\mathbb{P}_Z$ and $\mathbb{P}_{X+\varepsilon}$, i.e., the joint distributions of Z and $X+\varepsilon$ whose marginal distributions are $\mathbb{P}_Z$ and $\mathbb{P}_{X+\varepsilon}$. It is known that (64) is a metric for convergence in distribution among distributions with finite second moments, which means that it satisfies the axioms of distance and that $W_2(\mu, \nu_k) \to 0$ if and only if $\nu_k \xrightarrow{d} \mu$. Then, because (63) holds:

$$W_2(\mathbb{P}_Z, \mathbb{P}_{X+\varepsilon}) = 0.$$

Based on this result, the aim of Arellano and Bonhomme (2021) is to find $\mathbb{P}_X$ that minimizes (64). They obtain pseudo-observations of X by minimizing the sample version of (64).

I apply their approach in the panel data setting to obtain pseudo-observations of the permanent income. I assume that the transitory income process, ε_{it}, follows an i.i.d. (over i and t) zero-mean Normal distribution whose variance is equal to the variance estimate of the transitory income in Guvenen (2009). I then simulate $K = 200$ i.i.d. draws of the transitory income process:

$$\varepsilon_k = (\varepsilon_{k1}, \ldots, \varepsilon_{kT}) \in \mathbb{R}^T, \quad k = 1, \ldots, K.$$

Then, given the initial values of pseudo-observations of the permanent income defined by

$$\tilde{Y}_i = (\tilde{Y}_{i1}, \ldots, \tilde{Y}_{iT}) \in \mathbb{R}^T, \quad i = 1, \ldots, N,$$

I calculate the *synthetic* income data by calculating:

$$\hat{Y}_{ik} = \tilde{Y}_i + \varepsilon_k \in \mathbb{R}^T, \quad i = 1, \ldots, N, \quad k = 1, \ldots, K, \tag{65}$$

giving synthetic income data of size NK. Note that (65) computes a convolution of the

[10] They also consider a more general case of multivariate factor models.

permanent and the transitory income processes, because the empirical distribution of $\{\hat{Y}_{ik}\}$ is equal to the convolution of the empirical distribution of $\{\tilde{Y}_i\}$ and the empirical distribution of $\{\varepsilon_k\}$.

I then compare $\{\hat{Y}_{ik}\}$ with the observed income data in the PSID dataset, denoted by $Y_i = (Y_{i1}, \ldots, Y_{iT}) \in \mathbb{R}^T$, $i = 1, \ldots, N$. Let $\hat{P}_Y$ and $\hat{P}_{\hat{Y}}$ be the empirical distributions of $\{Y_i\}$ and $\{\hat{Y}_{ik}\}$, respectively. Then the (squared) second-order Wasserstein distance between the synthetic and the observed data is given by:

$$W_2^2(\hat{P}_Y, \hat{P}_{\hat{Y}}) = \min_{0 \leq p_{ijk} \leq 1} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{K} p_{ijk} ||Y_i - \hat{Y}_{jk}||^2$$

$$\text{subject to} \qquad \sum_{i=1}^{N} p_{ijk} = 1, \quad \sum_{j=1}^{N} \sum_{k=1}^{K} p_{ijk} = 1,$$

which is the sample version of (64). I then obtain pseudo-observations of the permanent income, $\tilde{Y}_i = (\tilde{Y}_{i1}, \ldots, \tilde{Y}_{iT})$, $i = 1, \ldots, N$, by:

$$\{\tilde{Y}_i\} = \underset{\tilde{Y}_1, \ldots, \tilde{Y}_N}{\operatorname{argmin}} W_2^2(\hat{P}_Y, \hat{P}_{\hat{Y}}),$$

which can be shown to be a convex optimization problem. This gives a dataset with $N = 800$ individuals and $T = 15$ waves.

This de-noising procedure does not qualitatively affect the estimation results in Section 8.4 of the main text. Tables 7 and 8 present confidence intervals for $\mathbb{E}(\rho_i)$, $\mathrm{Var}(\rho_i)$ and the CDF of ρ_i, computed without the de-noising step. The confidence intervals in Tables 7 and 8 are qualitatively similar to those in Tables 3 and 4 in the main text. For example, the upper confidence limit of $\mathbb{E}(\rho_i)$ is significantly less than 1, the lower confidence limit of $\mathrm{Var}(\rho_i)$ is strictly positive for the RIP process, and the CDF of ρ_i has confidence limits away from 0 and 1.

	$\mathbb{E}(\rho_i)$	$\mathrm{Var}(\rho_i)$
RIP process	[0.415, 0.652]	[0.073, 0.235]
HIP process	[0.262, 0.692]	[0.000, 0.659]

Table 5: Confidence intervals for $\mathbb{E}(\rho_i)$ and $\mathrm{Var}(\rho_i)$, computed without removal of individuals with small variations in their reported incomes. The nominal coverage probability is 0.9.

$\mathbb{P}(\rho_i \leq r)$	RIP process	HIP process
$r = 0.0$	[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 6: Confidence intervals for $\mathbb{P}(\rho_i \leq r)$, computed without removal of individuals with small variations in their reported incomes. The nominal coverage probability is 0.9.

	$\mathbb{E}(\rho_i)$	$\mathrm{Var}(\rho_i)$
RIP process	[0.451, 0.615]	[0.050, 0.293]
HIP process	[0.242, 0.633]	[0.000, 0.700]

Table 7: Confidence intervals for $\mathbb{E}(\rho_i)$ and $\mathrm{Var}(\rho_i)$, computed without the de-noising step and with the removal of individuals with small variations in their reported incomes. The nominal coverage probability is 0.9.

$\mathbb{P}(\rho_i \leq r)$	RIP process	HIP process
$r = 0.0$	[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 8: Confidence intervals for $\mathbb{P}(\rho_i \leq r)$, computed without the de-noising step and with the removal of individuals with small variations in their reported incomes. The nominal coverage probability is 0.9.