Estimating Latent-Variable Panel Data Models Using Parameter-Expanded SEM Methods *

Siqi Wei[†]
IE University

Work in progress, February 26, 2023
Click here for the latest version

Abstract

The Expectation-Maximization (EM) algorithm is a popular tool for estimating models with latent variables. In complex models, simulated versions such as stochastic EM, are often implemented to overcome the difficulties in computing expectations analytically. A drawback of the EM algorithm and its variants is the slow convergence in some cases, especially when the models contain relatively large number of latent variables. Liu et al., 1998 proposed a parameter-expanded algorithm (PX-EM) to speed up convergence. This paper explores the potential of parameter expansion ideas for estimating latent-variable panel models using the stochastic EM algorithm. I develop PX-SEM methods for three types of panel data models: 1) dynamic factor models, 2) binary choice models with individual effects and persistent shocks, and 3) persistent-transitory dynamic quantile processes. I find that PX-SEM could greatly speed up convergence.

JEL Codes: C13, C33, C63.

Keywords: Stochastic EM, Parameter-expansion, Dynamic factor model, Discrete choice model, Dynamic quantile regression, Latent variables.

*I am grateful to Manuel Arellano for his invaluable support and advice. I would also like to thank Martin Almuzara, Dmitry Arkhangelsky, Orazio Attanasio, Richard Blundell, Stéphane Bonhomme, Micole De Vera, Pedro Mira, Josep Pijoan-Mas, Enrique Sentana Liyang Sun for their great comments and discussions. Funding from Spain's Ministerio de Economía, Industria y Competitividad (BES-2017-082506) and María de Maeztu Programme for Units of Excellence in R&D (MDM-2016-0684) is gratefully acknowledged. All errors are my sole responsibility.

[†]IE University. Email: siqi.wei@ie.edu. Website: sites.google.com/cemfi.edu.es/siqiwei/

1 Introduction

The Expectation-Maximization (EM) algorithm proposed by Dempster et al., 1977 is a useful tool for empirical models with latent variables for obtaining maximum-likelihood estimates. Starting from an initial guess of parameters, the algorithm iterates between an E-step, which computes the conditional mean of certain functions of latent variables given observables, and an M-step, which solves the likelihood-based optimization problem and updates parameters until the convergence to the maximum of the likelihood. The EM algorithm has also been extended to introduce GMM estimation in the M-step (Arcidiacono and Jones, 2003). However, in complicated models where computing the E-step analytically is infeasible, simulated versions of the EM algorithm are often implemented. A prominent example is the stochastic expectation-maximization (SEM) algorithm (Diebolt and Celeux, 1993). In this case, the task of the E-step becomes drawing latent variables from the posterior distribution given observables, whereas the M-step becomes updating parameters as if the draws were observables.[1] Starting from an initial guess, one needs to iterate between two steps until the convergence of the estimates to the stationary distribution. The method could greatly simplify the implementation because the M-step optimization under pseudo-complete data is usually much easier.

A drawback of the EM algorithm and its variants is the slow convergence in some situations, especially when the models contain multiple latent variables over multiple periods, as in many panel data models, or when the initial guesses are not good. Indeed, as it is usually hard to know whether the initial guesses are good or not and to prevent the series converge to some local maximum, one strategy that researchers use is to run from a large amount of initial guesses and select based on some criteria such as likelihood. As a consequence, the slow convergence issue becomes even more prominent. Recent work has looked at alternative samplers for the latent variables when performing the E step. In contrast, we will focus on M step and try to improve the convergence rate especially for nonlinear panel data models.

To do so, this paper combines the parameter-expansion technique studied in Liu et al., 1998 with the SEM algorithm and develops PX-SEM algorithms for three types of nonlinear panel data models: 1) dynamic factor models, 2) binary choice models with individual effects, persistent and transitory components, and 3) persistent-transitory

[1] Arellano and Bonhomme, 2016 extends SEM algorithm by replacing the likelihood-based M-step by quantile regressions.

dynamic quantile processes.

Similarly, the PX-SEM algorithm also consists of two steps. The E-step is the same as in the standard SEM algorithm. In contrast, the M-step estimator is replaced by a more robust one, taking into account that the E-step draws under parameter values far from the optimum could violate model assumptions. Specifically, the M step involves: 1) expanding the original model (O model) to a larger one (L model), 2) estimating the L model, and 3) reducing to O model space, which all together aims to exploit extra information from the latent-variable model assumptions.

To implement the PX-SEM algorithm, one needs to develop a suitable expanded L model with auxiliary parameters and a reduction function. L model needs to satisfy two restrictions. First, there exists some value of auxiliary parameters such that L model equals O model. Second, there exists the reduction function, which is a mapping from L model parameter space to O model parameter space, such that the likelihood of observables is preserved. Everything else the same, a more flexible L model should not increase the number of iterations needed for convergence, but it might cause extra execution time for each iteration, and thus leads to longer computing time. Therefore, there is a tradeoff between the flexibility of the L model and additional complications in L model estimation and reduction.

Taking the discussion above into account, this paper develops procedures to use PX-SEM algorithms for three types of panel data models, the dynamic factor models, the discrete choice models, and persistent-transitory dynamic quantile processes. The focus on panel data models is expected to make the PX-SEM implementation more challenging but also more fruitful. On the one hand, panel data models are widely used in applied work. On the other hand, panel data models allow for more latent variables such as individual effects, and persistent and transitory components over multiple periods, which worsens the convergence issues of the SEM algorithm, but increases considerably the potential benefits of PX-SEM.

For all applications, we expand the O model linearly by allowing for a non-zero correlation between variables that are assumed to be non-correlated. Additionally, in some of the applications, we choose the L model such that the reduction function is simply identity mapping. Therefore, the only task left in M-step is to estimate the L model.

Finally, by doing simulation, we find that the PX-SEM algorithms significantly improve

the algorithmic efficiency compared to the standard SEM algorithm. A general lesson that we learn is that even without expanding the O model, from the perspective of reducing convergence time, we should consider estimators that require less latent variable information other than MLE.

Literature and contribution. This paper belongs to expanding literature that considers the application of the EM algorithm (Dempster et al., 1977) and its variants in estimating latent variable models (Diebolt and Celeux, 1993; Arcidiacono and Jones, 2003; Arellano and Bonhomme, 2016; Liu et al., 1998). This paper contributes to this literature in two ways. First, I combine the parameter expansion idea with the stochastic EM algorithm and discuss both likelihood-based and moments-based M-step estimators.[2] Additionally, we propose ways to implement PX-SEM algorithms for three types of nonlinear panel data models that are widely used in applied work: 1) dynamic factor models, 2) discrete choice models, and 3) persistent-transitory quantile models. In simulations, we show that PX-SEM could significantly reduce the convergence time.

Organization. The paper proceeds as follows. In Section 2, I illustrate the difference between the standard stochastic EM algorithm and the parameter-expanded stochastic EM algorithm using a simple toy model. Section 3 defines the PX-SEM algorithm and discuss its statistical properties. Next, I develop PX-SEM methods for three types of nonlinear latent-variable panel data models. In Sections 4 to 6, I propose the PX-SEM algorithms for dynamic factor models, discrete choice models, and persistent-transitory dynamic quantile processes, respectively. Finally, Section 7 concludes.

2 Toy Model

In this section, we will compare the standard stochastic EM (SEM) algorithm with the parameter-expanded stochastic EM (PX-SEM) algorithm based on a simple toy model. In addition to showing the difference between the two methods, we will also explain intuitively why PX-SEM might speed up the convergence.

Consider the following model that we want to estimate:

O Model:

$$y_i = y_i^* + \epsilon_i, \quad \begin{pmatrix} y_i^* \\ \epsilon_i \end{pmatrix} \sim N\big(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix}\big)$$

[2] Liu et al., 1998 is based on the standard EM algorithm. Liu and Wu, 1999 applies the parameter expansion technique to Bayesian inference by combining with the data augmentation algorithm; Lavielle and Meza, 2007 combines the parameter expansion technique with Monte Carlo EM.

where $y_1, ..., y_N$ are observed outcomes and $y_1^*, ..., y_N^*$ are latent variables whose distribution is of interest. The only unknown parameter is the standard deviation σ. In fact, this model is simple enough for us to write down the closed-form of the log-likelihood function and the maximum likelihood estimator, such that there is no need to implement SEM or PX-SEM algorithms. However, we will use this model to illustrate two algorithms, respectively.

SEM algorithm. To implement SEM algorithm, we need to start with a guess of unknown parameter $\hat{\sigma}^{(0)}$, and then iterate the following two steps on $s = 0, 1, 2, ..., S$ until the convergence of $\hat{\sigma}^{(s)}$ to the stationary distribution:

1. Stochastic E step: Draw y_i^* from posterior distribution $f_O(y_i^*|y_i; \hat{\sigma}^{(s)})$
2. M step: Update parameters by computing $\hat{\sigma}^{(s+1)} = \arg\max_\sigma \sum_i l_O(\sigma; y_i^*, y_i)$, that is $\hat{\sigma}^{(s+1)} = \widehat{\text{std}}(y_i^*)$

where $f_O(\cdot; \sigma)$ is the density function of O Model, and $l_O(\cdot; y^*, y)$ is the log-likelihood function of pseudo-complete data. The final estimator is the average of last S^0 iterations $\hat{\sigma} = \frac{1}{S^0} \sum_{S-S^0+1}^{S} \hat{\sigma}^{(s)}$

EM algorithm works because it improves the observed-data likelihood in each iteration:

$$\log f_O(y_i; \hat{\sigma}^{(s+1)}) - \log f_O(y_i; \hat{\sigma}^{(s)}) \geq Q(\hat{\sigma}^{(s+1)}|\hat{\sigma}^{(s)}) - Q(\hat{\sigma}^{(s)}|\hat{\sigma}^{(s)}) \geq 0$$

where $Q(\hat{\sigma}^{(s+1)}|\hat{\sigma}^{(s)}) = \int \log f_O(y_i, y_i^*; \hat{\sigma}^{(s+1)}) f_O(y_i^*|y_i; \hat{\sigma}^{(s)}) dy_i^*$

PX-SEM algorithm. Now we introduce the PX-SEM algorithm. Similar to the SEM algorithm, PX-SEM also consists of an E-step where we draw latent variables and an M-step where we update parameters. The E-step is the same as in the SEM algorithm, whereas in the M-step, PX-SEM requires 1) expanding the original model, 2) estimating the expanded one, and 3) reducing to the original model space to obtain the estimator. For notation simplicity, we refer to the expanded larger model as the L model relative to the original model (O model).

For this toy model, we propose the following L model:

L Model:

$$y_i = y_i^* + \epsilon_i$$

where $\begin{pmatrix} y_i^* \\ \epsilon_i \end{pmatrix} \sim N\left(0, K \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix} K'\right)$, $K = \begin{pmatrix} k & 0 \\ 1-k & 1 \end{pmatrix}$

In addition to σ, the L model also contains an auxiliary parameter k. It is easy to show that when $k = 1$, the two models coincide, that is $f_O(y_i^*, y_i; \sigma) = f_L(y_i^*, y_i; k = 1, \sigma)$;

and when $k \neq 1, 0$, L model expands the O model by allowing for a non-zero correlation between y_i^* and ϵ_i, as $\text{cov}(y_i^*, \epsilon_i) = k(1-k)\sigma^2$.

Now we implement the PX-SEM algorithm. Starting with a guess of unknown parameter $\hat{\sigma}^{(0)}$, we iterate the following two steps on $s = 0, 1, 2, ..., S$ until the convergence of $\hat{\sigma}^{(s)}$ to the stationary distribution:

1. Stochastic E step: Draw y_i^* from posterior distribution $f_O(y_i^*|y_i; \hat{\sigma}^{(s)})$
2. PX-M step: Update parameters by
   (a) Estimate the L model: computing $(\hat{\sigma}_L^{(s+1)}, \hat{k}_L) = \arg\max_{\sigma,k} \sum_i l_L(\sigma, k; y_i^*, y_i)$, that is $\hat{k}_L = \frac{\widehat{\text{var}}(y_i^*)}{\widehat{\text{cov}}(y_i^*, y_i)}$, $\hat{\sigma}_L^{(s+1)} = \frac{\widehat{\text{std}}(y_i^*)}{|\hat{k}_L|}$
   (b) Obtain $\hat{\sigma}^{(s+1)}$ by mapping the L model to the O model space while keeping $f_O(y_i; \hat{\sigma}^{(s+1)}) = f_L(y_i; \hat{k}_L, \hat{\sigma}_L^{(s+1)})$, that is $\hat{\sigma}^{(s+1)} = \hat{\sigma}_L^{(s+1)}$

The final estimator is the average of last S^0 iterations:$\hat{\sigma} = \frac{1}{S^0} \sum_{S-S^0+1}^{S} \hat{\sigma}^{(s)}$.

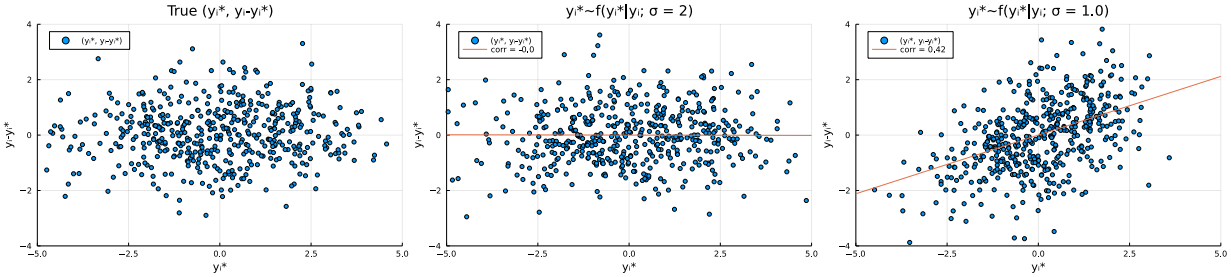
As described above, the two methods share the same E step, and the only difference is in the estimators of the M step: the PX-SEM estimator is adjusted by $\frac{1}{\hat{k}_L}$. Figure 1 explains what the PX-SEM and its auxiliary parameter k do. Specifically, Figure 1a is the scatter plot of simulations from the O model with a true value of $\sigma = 2$. The X-axis and Y-axis display y_i^* and $\epsilon_i = y_i - y_i^*$, respectively. As expected, we do not observe significant correlations between the sample y_i^* and ϵ_i. Remember, we assume that y_i^* is latent, and only y_i is observed. Next, given y_i, we conduct E-step and compare the y_i^* draws under different guesses of the value of σ.

Figure 1b is the scatter plot of $(\hat{y}_i^*, y_i - \hat{y}_i^*)$ where $\hat{y}_i^*$ is the E-step draws under the true value, that is $\hat{y}_i^* \sim f_O(y_i^*|y_i; \sigma = 2)$, and Figure 1c is the scatter plot of $(\hat{y}_i^*, y_i - \hat{y}_i^*)$ where $\hat{y}_i^*$ is the E-step draws under a wrong value $\sigma = 1$, that is $\hat{y}_i^* \sim f_O(y_i^*|y_i; \sigma = 1)$. Figure 1c presents a significant positive correlation between $\hat{y}_i^*$ and $y_i - \hat{y}_i^*$, which should be zero by assumption. We generate this "false" positive correlation because the draws are taken under the "constraints" that 1) the variance of y_i^* should not be far away from 1, and 2) the variance of $y_i - y_i^*$ should not be far away from 1.[3]

[3]The "constraints" are from the distribution from which we draw y_i^*: $f(y_i^*|y_i; \sigma = 1) \propto \phi(y_i^*)\phi(y_i - y_i^*)$

Figure 1: Data and draws of E-step under different guess of σ

(a) True $(y_i^*, y_i - y_i^*)$ (b) Draws from E-step given $\sigma^{(s)} = \sigma_{true}$, $(y_i^*, y_i - y_i^*)$ (c) Draws from E-step given $\sigma^{(s)} \neq \sigma_{true}$, $(y_i^*, y_i - y_i^*)$



In the case of Figure 1b, both M-step estimators are consistent, and the SEM one is more efficient due to the correct constraints ($k = 1$). However, In the case of Figure 1c, the M-step of SEM ignores the violation of the zero-correlation assumption. As a consequence, the estimator $\hat{\text{std}}(y_i^*)$ is no more consistent. In contrast, PX-SEM takes into account the "false" correlation by adding parameter k and the L model estimator is still consistent. To put it another way, in this case, the M-step estimator of the SEM algorithm is the pseudo MLE, whereas the PX-SEM is the MLE, which leads to a larger increment of likelihood changes (of complete data y_i and y_i^*). And finally, by reducing to the O model space keeping the likelihood of observed data y_i unchanged, we preserve the "gains" in the complete data likelihood.

There are potentially a lot of ways of expanding the original model, and considerations include how easy to estimate the L model and reduce it to O model. Section 3 has further discussion, and Appendix B compares different L models as well as M-step estimators using the toy model example. Yet it is worth noting that if we propose the following L model: $y_i = y_i^* + \epsilon_i$, where $\begin{pmatrix} y_i^* \\ \epsilon_i \end{pmatrix} \sim N\left(0, K \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix} K'\right)$, $K = \begin{pmatrix} k_1 & k_2 \\ 0 & k_3 \end{pmatrix}$, subject to $CK \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix} K'C' = C \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix} C'$, $C = \begin{pmatrix} 1 & 1 \end{pmatrix}$, which jointly with the normality assumption guarantees that $f_O(y_i; \sigma) = f_L(y_i; \sigma, K)$, then the PX-SEM estimator is the GMM estimator: $\hat{\sigma} = \widehat{\text{var}}(y_i) - 1$.

3 Parameter Expansion Stochastic EM Algorithm

In this section, I will first define the PX-SEM algorithm and explain the implementation steps in a general way. Next, I will discuss the statistical properties and the reason why it could improve the algorithmic efficiency.

3.1 Definition of PX-SEM algorithm

Setup. Let $\{Y_i, X_i, Y_i^*\}$ for $i = 1 : N$ be i.i.d. random variables from the O Model distribution $f_O(Y_i|X_i;\theta) = \int_{Y_i^*} f_O(Y_i, Y_i^*|X_i;\theta)dY_i^*$, where $W_i \equiv [Y_i; X_i']'$ is the observable set, Y_i^* is the latent-variable set, and θ is the unknown parameter set to be estimated.The true value, $\overline{\theta}$, satisfies the equation $E(\Psi_O(Y_i^*, W_i;\overline{\theta})) = 0$, where $\Psi_O(\cdot)$ represents the score function of the complete O model in the case of likelihood-based PX-SEM algorithm and moment restrictions in the case of moment-based one. We can easily show that the true value $\overline{\theta}$ also satisfies the equation

$$E\Big(\int \Psi_O(Y_i^*, W_i;\overline{\theta}) f_O(Y_i^*|W_i;\overline{\theta})dY_i^*\Big) = 0 \tag{1}$$

Denote the expanded model, L Model, by $f_L(Y_i|X_i;\theta,K) = \int_{Y_i^*} f_L(Y_i, Y_i^*|X_i;\theta,K)dY_i^*$, where K represents for all auxiliary parameters. The expanded L model needs to satisfy two conditions: (1) L model nests O model: $\exists\ K_0$ such that $f_O(Y_i, Y_i^*|X_i;\theta) = f_L(Y_i, Y_i^*|X_i;\theta,K_0)$, $\forall\theta$, and (2) There is a mapping, the reduction function, from the L Model space to O Model space $\theta = R(\theta_L, K)$ such that the observed data likelihood is preserved $f_O(Y_i|X_i;R(\theta_L,K)) = f_L(Y_i|X_i;\theta_L,K)$, $\forall\theta_L, K$.[4]

Function $\Psi_L^\theta(\cdot)$ represents the score of the L model with respect to θ in the case of likelihood-based PX-SEM algorithm and the same moment restrictions as $\Psi_O(\cdot)$ in the case of moment-based one. Under condition (1), we have $\Psi_L^\theta(Y_i^*, W_i;\theta,K_0) = \Psi_O(Y_i^*, W_i;\theta)$, and thus $E(\Psi_L^\theta(Y_i^*, W_i;\overline{\theta},K_0)) = 0$. Additionally, assume that there exist moment restrictions $\Psi_L^K(\cdot)$ such that K is identified when we observe Y_i^*, that is $E(\Psi_L^K(Y_i^*, W_i;\overline{\theta},K_0)) = 0$, then we have $E(\Psi_L(Y_i^*, W_i;\overline{\theta},K_0)) = 0$, where $\Psi_L(\cdot) = [\Psi_L^\theta(\cdot);\Psi_L^K(\cdot)]$. Equivalently, we have

$$E\Big(\int \Psi_L(Y_i^*, W_i;\overline{\theta},K_0) f_O(Y_i^*|W_i;R(\overline{\theta},K_0))dY_i^*\Big) = 0 \tag{2}$$

Definition of PX-SEM algorithm. Before we introduce the general steps of the PX-SEM algorithm, for comparison, let us have a look at the SEM algorithm. SEM is an iterative algorithm where in the E step we draw latent variables Y_i^* from posterior distribution $f_O(Y_i^*|W_i;\hat{\theta}^{(s)})$ under parameter guess $\hat{\theta}^{(s)}$, and in the M step update parameters to $\hat{\theta}^{(s+1)}$, that is $\sum_i\left(\Psi_O(Y_i^*, W_i;\hat{\theta}^{(s+1)})\right) = 0$. The stochastic version differs from the original EM algorithm because we replace the integral by the latent draws in equation (C1).[5]

[4] We know reduction function should satisfy $R(\theta,K_0) = \theta$.

[5] The Monte Carlo EM uses many simulations to approximate the conditional expectation whereas the SEM uses only one or few in each iteration (Wei and Tanner, 1990; Nielsen, 2000).

In contrast, PX-SEM algorithm proposes the iterations which are better linked to equation (C2): we still draw latent variables Y_i^* from posterior distribution $f_O(Y_i^*|W_i;\hat{\theta}^{(s)})$ under parameter guess $\hat{\theta}^{(s)}$, but we use the expanded model to update to $\hat{\theta}^{(s+1)}$.

The general steps are as follows: starting with a guess of unknown parameter $\hat{\theta}^{(0)}$, we iterate the following two steps on $s = 0, 1, 2, ..., S$ until the convergence of $\hat{\theta}^{(s)}$ to the stationary distribution:

1. Stochastic E step: Draw Y_i^* from posterior distribution $f_O(Y_i^*|W_i;\hat{\theta}^{(s)})$
2. PX-M step: Update parameters by
   (a) Estimate L model: $\sum_i \Psi_L(Y_i^*, W_i; \hat{\theta}_L^{(s+1)}, \hat{K}^{(s+1)}) = 0$
   (b) Reduction: $\hat{\theta}^{(s+1)} = R(\hat{\theta}_L, \hat{K})$ subject to $f_O(Y_i|X_i;\hat{\theta}^{(s+1)}) = f_L(Y_i|X_i;\hat{\theta}_L,\hat{K})$

In practice, one of the challenges in choosing appropriate L model is to figure out the associated reduction function. A strategy that this paper takes in most of the following applications is to expand the model in a specific way such that the reduction function is simply $\theta = R(\theta, K)$. In more detail, in the following sections where I develop PX-SEM algorithms for discrete choice models and quantile models, I *choose* L models where auxiliary parameter K does not affect the observed data likelihood:

$$f_O(Y_i|X_i;\theta_L) = f_L(Y_i|X_i;\theta_L, K) \tag{3}$$

The advantage the extra restriction brings to us is that it implies $R(\theta_L, K) = \theta_L$, and therefore the procedure of PX-SEM algorithm can be simplified as:

1. Stochastic E step: Draw Y_i^* from posterior distribution $f_O(Y_i^*|X_i, Y_i;\hat{\theta}^{(s)})$
2. PX-M step: Update parameters by estimating the L model:

$$\sum_i \Psi_L(Y_i^*, W_i; \hat{\theta}^{(s+1)}, \hat{K}) = 0$$

By comparing the M-steps of SEM algorithm with the PX-SEM algorithm, we can see that the estimator of the SEM M-step is a restricted version of PX-SEM M-Step estimator under the restriction of $K = K_0$. When the draws Y_i^* are taken under a guess $\hat{\theta}^{(s)}$ that is close enough to the true value, intuitively, we expect the SEM estimator to be more efficient given the correct restriction $K = K_0$. In this case, the PX-SEM estimator is still consistent based on equation (C2) which becomes $E\left(\int \Psi_L(Y_i^*, W_i; \overline{\theta}, K_0) f_O(Y_i^*|X_i, Y_i; \overline{\theta}) dY_i^*\right) = 0$ given the extra restriction (3). However, under the guess $\hat{\theta}^{(s)}$ which is far enough such that the draws Y_i^* violate some model assumptions, we would expect the PXSEM estimator

to be more "robust" given the extra flexibility in K. Correspondingly, as we will show in the following subsection, the likelihood-based PX-M-step could achieve a higher pseudo complete data likelihood improvement, which further leads to a higher observed-data likelihood improvement in each iteration compared to the SEM M-step.

3.2 Statistical properties

In this subsection, we will first show that likelihood-based parameter expanded EM algorithm increases the log-likelihood of the observed-data model at each iteration based on the work of Liu et al., 1998. Then combining with Nielsen, 2000 and Arellano and Bonhomme, 2016, we give conditions under which the stochastic version, PX-SEM, on average dominates the SEM in global rate of convergence. Finally, we discuss the asymptotic properties of the PX-SEM estimator.

Convergence. When the M step is likelihood-based, that is when the $\Psi_L(Y_i^*, W_i; \theta, K) = [\frac{\partial}{\partial\theta} \ln f_L(Y_i, Y_i^*|X_i; \theta, K); \frac{\partial}{\partial K} \ln f_L(Y_i, Y_i^*|X_i; \theta, K)]$, Liu et al., 1998 shows that the parameter expanded EM algorithm, just like the original EM algorithm, increases the loglikelihood of the observed-data model at each iteration. Here we discuss this briefly.

It can be easily shown that:

$$\log f_O(Y_i|X_i; \hat{\theta}^{(s+1)}) - \log f_O(Y_i|X_i; \hat{\theta}^{(s)}) = \log f_L(Y_i|X_i; \hat{\theta}_L, \hat{K}) - \log f_L(Y_i|X_i; \hat{\theta}^{(s)}, K_0)$$

The equation holds because of both condition (1), that is L model nests O model and $\exists K_0$, $f_O(Y_i|X_i; \hat{\theta}^{(s)}) = f_O(Y_i|X_i; \hat{\theta}^{(s)}, K_0)$, and condition (2), that is the reduction function exists — so by construction $f_O(Y_i|X_i; \hat{\theta}^{(s+1)}) = f_L(Y_i|X_i; \hat{\theta}_L, \hat{K})$.

Then, given Gibbs' inequality, we have

$$\sum_i \log f_L(Y_i|X_i; \hat{\theta}_L, \hat{K}) - \sum_i \log f_L(Y_i|X_i; \hat{\theta}^{(s)}, K_0) \geq Q(\hat{\theta}_L, \hat{K}|\hat{\theta}^{(s)}, K_0) - Q(\hat{\theta}^{(s)}, K_0|\hat{\theta}^{(s)}, K_0)$$

where $Q(\hat{\theta}_L, \hat{K}|\hat{\theta}^{(s)}, K_0) = \sum_i \int \log f_L(Y_i, Y_i^*|X_i; \hat{\theta}_L, \hat{K}) f_L(Y_i^*|Y_i, X_i; \hat{\theta}^{(s)}, K_0) dY_i^*$.

Finally, using definition of $\hat{\theta}_L$, which is $\hat{\theta}_L, \hat{K} = \arg\max_{\theta,K} Q(\theta, K|\hat{\theta}^{(s)}, K_0)$, we can prove

$$\sum_i \log f_O(Y_i|X_i; \hat{\theta}^{(s+1)}) - \sum_i \log f_O(Y_i|X_i; \hat{\theta}^{(s)}) \geq 0$$

The inequality above says that the PX-EM improves the observed-data likelihood in each iteration. Moreover, the L model being more flexible and nesting the O model implies the following inequality:

$$Q(\hat{\theta}_L, \hat{K}|\hat{\theta}^{(s)}, K_0) - Q(\hat{\theta}^{(s)}, K_0|\hat{\theta}^{(s)}, K_0) \geq Q(\hat{\theta}_{SEM}^{(s+1)}, K_0|\hat{\theta}^{(s)}, K_0) - Q(\hat{\theta}^{(s)}, K_0|\hat{\theta}^{(s)}, K_0)$$

where $\hat{\theta}_{SEM}^{(s)} = \arg\max_\theta \sum_i \int \log f_O(Y_i, Y_i^*|X_i;\theta) f_L(Y_i^*|Y_i, X_i; \hat{\theta}^{(s)}, K_0) dY_i^*$.

Therefore, the parameter expansion technique can be intuitively interpreted as a way to improve the lower bound of the loglikelihood increment.

Convergence speed. In Appendix C.3, we discuss in detail the conditions under which parameter expansion technique speeds up the convergence. Importantly, one implication is that with likelihood-based M step, PX-SEM on average dominates the SEM in computational efficiency. Specifically, defining $\hat{\theta}$ as the MLE, we can write the dynamics of SEM iterations $\hat{\theta}_{SEM}^{(s)}$ as follows:

$$(\hat{\theta}_{SEM}^{(s+1)} - \hat{\theta}) = (I - V_{SEM})(\hat{\theta}_{SEM}^{(s)} - \hat{\theta}) + A_{SEM}\epsilon_\theta^{(s)} + o_p(N^{-(1/2)})$$

and the dynamics of PX-SEM iterations $\hat{\theta}^{(s)}$ as follows:

$$(\hat{\theta}^{(s+1)} - \hat{\theta}) = (I - V_{PX})(\hat{\theta}^{(s)} - \hat{\theta}) + A_{PX}\epsilon^{(s)} + o_p(N^{-(1/2)})$$

where the expression of V_{SEM}, V_{PX}, A_{SEM}, and A_{px} are given in Appendix C.3. We show in the appendix that the smallest eigenvalue of V_{PX}, which is the global speed, is at least as large as the smallest eigenvalue of V_{SEM}. Therefore, the PX-SEM on average exhibits a higher global convergence speed than SEM.

Asymptotic properties. Nielsen, 2000 studies the statistical properties of likelihood-based SEM algorithm. Specifically, the paper charaterizes the asymptotic distribution of $\sqrt{N}(\hat{\theta}^{(s)} - \overline{\theta})$, when sample size tend to infinity, where $\overline{\theta}$ is the true value of θ. Arellano and Bonhomme, 2016 expands the results by discussing the asymptotic properties for moment-based SEM algorithm. Based on these two papers, in Appendix C.2, we show that in case of convergence and s corresponds to a draw from the ergodic distribution of the Markov chain, then

$$\sqrt{N}(\hat{\theta}^{(s)} - \overline{\theta}) \xrightarrow{d} \mathcal{N}(0, \Sigma_1 + \Sigma_2^{-1}\Sigma_3\Sigma_2^{-1\prime})$$

where the expression of Σ_1, Σ_2, and Σ_3 are given in Appendix C.2.

When the M-step is moment-based, in general, convergence is not guaranteed. In case of convergence, the speed does not necessarily dominate the SEM algorithm. In fact, in Appendix B, we show an example where the moment-based M step combined with parameter expansion technique works worse than SEM at least for some initial guesses.

However, in practice, we might still want to use the moments-based PX-SEM estimator for at least two reasons. First, in some cases, GMM estimators are much easier to obtain, such as the quantile example that we will discuss in Section 6. We care about the total

amount of time to converge which depends not only on the number of iterations but also on the time spent in each iteration. Secondly, even if it is still feasible to obtain MLE of the O model, restricting ourselves to tractable ML estimators in the M step might limit the flexibility in constructing the L model, which has effects on the speed of convergence. For example, in Appendix B, we show an example when the moment-based PX-SEM with a more flexible L model outperforms the MLE-based PX-SEM with a less flexible L model in the toy model case.

It is worth emphasizing again that the contribution of this paper is twofold. First, we combine the parameter-expansion technique developed in Liu et al., 1998 with the stochastic EM algorithm. Moving towards the stochastic EM version allows us to deal with more complicated models, such as nonlinear panel data models, where computing E-step analytically as required in the original EM algorithm is not feasible, and where the benefit of PX-SEM is expected to be large due to a higher dimension of latent variables.

Second and more important, given that the PX-SEM algorithm itself does not speak of selection of L model nor detailed steps of estimation, the other contribution of this paper is to propose specific L models and estimation steps to implement PX-SEM for nonlinear panel data models. In the following three sections, we will discuss three examples: 1) dynamic factor models, 2) discrete choice models, and 3) quantile models.

4 Dynamic Factor Models

The first example we explain is dynamic factor models (Geweke, 1977). The appeal of this class of models is that it can explain variation across multiple dimensions using variation in fewer latent common factors. Applications include topics in macroeconomics and finance.(Bai et al., 2008; Stock and Watson, 2006, 2011). The specific O model we discuss is a single factor model, but the same approach to implementing PX-SEM algorithm can be applied to models with multiple latent factors.

O Model:

$$y_{it} = \lambda_i \nu_t + \epsilon_{it}$$

$$\nu_t = \nu_{t-1} + u_t$$

where $\epsilon_{it} \sim N(0, \sigma_i^2)$, $u_{it} \sim N(0, 1)$.

There is a latent common factor ν_t that follows a Gaussian random walk. We observe N different measures, y_i, $i = 1, ..., N$, over in total T periods, and each of them is associated

with a different factor loading λ_i. In this model, we also assume ϵ_{it} is independent across periods.[6] Denote the set of unknown parameters by $\theta \equiv (\lambda_1, ..., \lambda_N, \sigma_1, ..., \sigma_N)$.

SEM algorithm. For comparison, we first explain the procedure of SEM algorithm. Starting from a guess $\hat{\theta}^{(0)}$, we iterate between the E-step and M-step on $s = 0, 1, 2, ..., S$ until the convergence of $\hat{\theta}^{(s)}$ to the stationary distribution:

1. Stochastic E step: Draw ν from posterior distribution $f_O(\nu|y; \hat{\theta}^{(s)})$
2. M step: $\hat{\theta}^{(s+1)} = (\hat{\lambda}_1, ..., \hat{\lambda}_N, \hat{\sigma}_1, ..., \hat{\sigma}_N)$

$$\hat{\lambda}_i, \hat{\sigma}_i = \max_{\lambda_i, \sigma_i} \sum_t \left(\ln \phi(\frac{y_{it} - \lambda_i \nu_t}{\sigma_i}) - \ln \sigma_i \right)$$

PX-SEM algorithm. To implement PX-SEM algorithm, we need to build a proper L model. In this case, we propose a very simple L model.

L model:

$$y_{it} = \lambda_i \nu_t + \epsilon_{it}$$

$$\nu_t = \nu_{t-1} + u_t$$

where $\epsilon_{it} \sim N(0, \sigma_i^2)$ and $u_{it} \sim N(0, \mathbf{k}^2)$

We expand the O model by adding an auxiliary parameter $\mathbf{k}$ such that the variance of persistent shock u_t could be different from 1. It is easy to verify the L model satisfies the condition (1), since we could always take $\mathbf{k} = 1$ and then the two models coincide.

Related to condition (2), there exists reduction function $R(\lambda_1, ..., \lambda_N, \sigma_1, ..., \sigma_N, \mathbf{k}) = (\lambda_1 \mathbf{k}, ..., \lambda_N \mathbf{k}, \sigma_1, ..., \sigma_N)$ such that the likelihood of the observed data is kept the same $f_O(y; \lambda_1 \mathbf{k}, ..., \lambda_N \mathbf{k}, \sigma_1, ..., \sigma_N) = f_L(y; \lambda_1, ..., \lambda_N, \sigma_1, ..., \sigma_N, \mathbf{k})$.

With the specified L model, we can implement the PX-SEM algorithm with the following procedures: We start from a guess $\hat{\theta}^{(0)}$, and iterate between the following E-step and PX-M step on $s = 0, 1, ..., S$ until the convergence of $\hat{\theta}^{(s)}$ to the stationary distribution:

1. Stochastic E step: Draw ν from posterior distribution $f_O(\nu|y; \hat{\theta}^{(s)})$
2. PX-M step:
   (a) L model estimation: $\hat{\lambda}_L, \hat{\sigma}_L = (\hat{\lambda}_{L1}, ..., \hat{\lambda}_{LN}, \hat{\sigma}_{L1}, ..., \hat{\sigma}_{LN})$

$$\hat{\lambda}_{Li}, \hat{\sigma}_{Li} = \max_{\lambda_i, \sigma_i} \sum_t \left(\ln \phi(\frac{y_{it} - \lambda_i \nu_t}{\sigma_i}) - \ln \sigma_i \right)$$

$$\hat{\mathbf{k}} = \max_{\mathbf{k}} \sum_t \left(\ln \phi(\frac{\nu_t - \nu_{t-1}}{\mathbf{k}}) - \ln \mathbf{k} \right)$$

[6] The method can be easily adapted to models with 1) unknown persistence in ν_t process, 2) multiple latent factors, 3) ϵ_{it} following MA process, etc.

(b) Reduction: $\hat{\theta}^{(s+1)} = (\hat{\lambda}_L \hat{\mathbf{k}}, \hat{\sigma}_L)$

As indicated in the PX-M step, due to the separability of the log-likelihood function, the auxiliary parameter $\mathbf{k}$ is very easy to estimate. Compared to the SEM, the PX-SEM update $\hat{\theta}^{(s+1)}$ takes into account the potential violation from the assumption $\mathbf{k} = 1$ in the O model. When the guess $\hat{\theta}^{(s)}$ is close enough to the true value, then we should expect $\hat{\mathbf{k}}$ is close to 1, and thus the estimates of SEM and PXSEM steps are similar. When the guess $\hat{\theta}^{(s)}$ is far enough from the true value such that the draw ν in the E step violates the assumption $\mathbf{k} = 1$, the PX-SEM algorithm adjusts the estimates accordingly. For example, when the $\hat{\mathbf{k}}$ is larger than 1, this suggests to scale down the latent draws ν by $\mathbf{k}$ to have $\text{var}(\Delta\nu) = 1$, and to scale up λ_i by $\mathbf{k}$ to have the same loglikelihood of observed data.

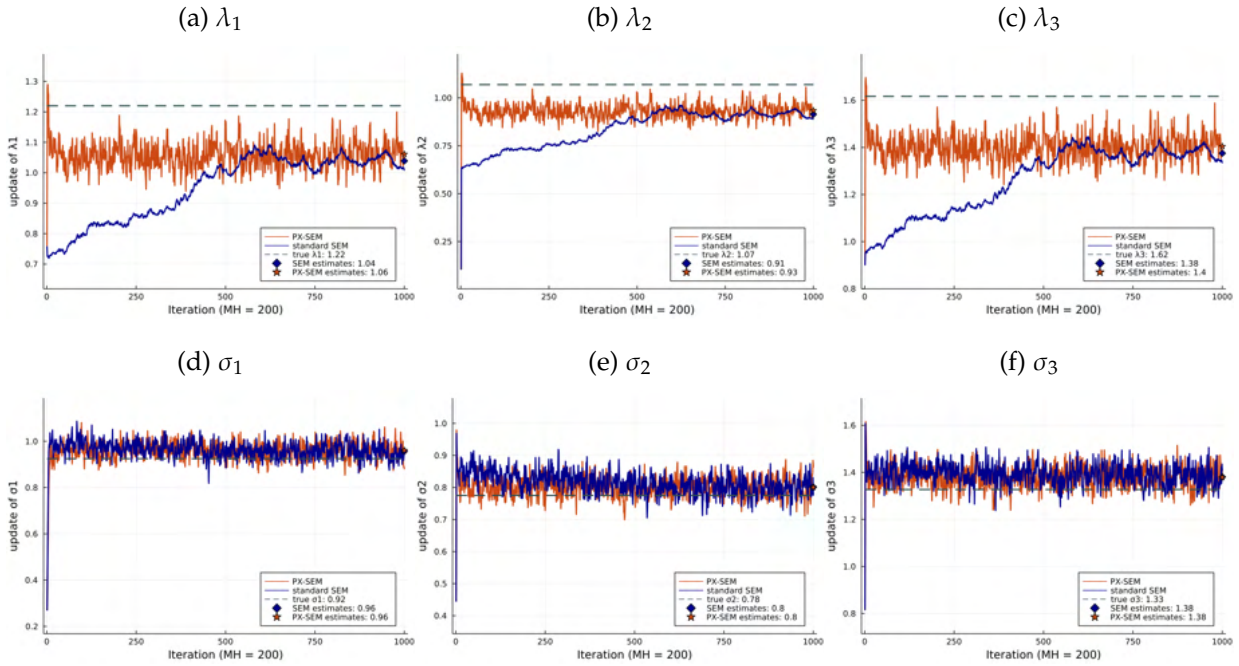
Simulation Results. Figure 2 presents the results based on one simulation of $N = 3$ and $T = 200$. The parameter values of the DGP are $\lambda = (1.22, 1.07, 1.62)$, and $\sigma = (0.92, 0.78, 1.33)$.

In each of the plots of Figure 2, the x-axis represents the number of iterations $s = 1, ..., 1000$, and y-axis represents the M-step update $\hat{\theta}^{(s)}$. The blue line depicts the SEM trajectory whereas the orange line depicts the PX-SEM trajectory. The horizontal green dash line represents for the true value. Starting from some randomly chosen initial guess $\hat{\theta}^{(0)}$, we see both the blue and the orange lines move towards the green dash line and become stable after some iterations. The average of last 250 iterations is taken as the final estimate.

Even though in the case of $\sigma's$, both methods converge almost immediately, we observe big difference in the case of $\lambda's$. Specifically, SEM updates do not seem to converge until 500 iterations, whereas the PX-SEM updates converge within 100 iterations. Nevertheless, after convergence, we observe larger variations among PX-SEM updates along iterations than the SEM updates. This is expected given that the SEM M-step is the MLE under correct restrictions. Our presumption is that in both cases, we average over large enough number of iterations after convergence. Whether PX-SEM should have more iterations due to larger variation, which might affect total computing time, is outside the scope of this paper.[7]

[7]We could always switch to SEM after PX-SEM estimators converge. Naturally, to what extent latent draws from the E-step violate the model assumptions (e.g., by comparing the O model and L model likelihood of pseudo-complete data) could be criteria for deciding when to switch.

Figure 2: SEM and PX-SEM iterations of $\theta^{(s)}$ from random initial guesses



Notes: Complete iterations of SEM (blue solid line), PX-SEM (orange solid line) based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond), PX-SEM estimates (orange star) are all based on the average of last 250 iterations. Random initial guess from lognormal distribution. $N = 3, T = 200$

Comment. The following model is equivalent to the proposed L model above, yet has different interpretation:

L Model

$$y_{it} = \lambda_i \nu_t + \epsilon_{it}$$

$$\begin{bmatrix} \nu \\ \epsilon_i \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{1}{\mathbf{k}} \times I_{T\times T} & 0_{T\times T} \\ \lambda_i(1-\frac{1}{\mathbf{k}}) \times I_{T\times T} & I_{T\times T} \end{bmatrix}}_{\mathbf{A_i}} \begin{bmatrix} \nu^* \\ \epsilon_i^* \end{bmatrix}$$

$$\nu_t^* = \nu_{t-1}^* + u_t, \quad \epsilon_{it}^* \sim N(0, \sigma_i^2), \quad u_{it} \sim N(0,1)$$

We add ν_t^* and ϵ_{it}^* which are assumed to have the same distributions as their counterparts in the O model. However, the L model extends the O model by allowing the latent draws from the E-step ν and ϵ_i to be results of an affine mapping of ν^* and ϵ_i^* through matrix $\mathbf{A_i}$. Related to condition (1), we could always have $\mathbf{k} = 1$, and thus $\mathbf{A_i} = I$, and L model is equal to O model. Moreover, matrix $\mathbf{A_i}$ satisfies the equation $C\mathbf{A_i} = C$ where $C = [I_{T\times T} \; I_{T\times T}]$, and thus $y_i = C[\nu' \; \epsilon_i']' = C[\nu^{*\prime} \; \epsilon_i^{*\prime}]'$. Therefore, relate to condition (2), auxiliary parameter $\mathbf{k}$ does not affect the observed data likelihood, which means the reduction function is $R(\theta, \mathbf{k}) = \theta$.

This L model is equivalent to the previous one in the sense that they have exactly the same estimators in the PX-M step for $\hat{\theta}^{(s+1)}$. Indeed, given the same E-step, the PX-M steps now becomes: $\widehat{\lambda_{Li}\mathbf{k}}, \hat{\sigma}_{Li} = \max_{\lambda_i \mathbf{k}, \sigma_i} \sum_t \left(\ln \phi(\frac{y_{it} - \lambda_i \mathbf{k} \nu_t}{\sigma_i}) - \ln \sigma_i \right)$, $\frac{1}{\hat{\mathbf{k}}} = \max_{\mathbf{k}} \sum_t \left(\ln \phi(\frac{\nu_t - \nu_{t-1}}{\mathbf{k}}) - \ln \mathbf{k} \right)$, and therefore $\hat{\theta}^{(s+1)} = (\hat{\lambda}_{L1}, ..., \hat{\lambda}_{LN}, \hat{\sigma}_{L1}, ..., \hat{\sigma}_{LN})$, given $R(\theta, \mathbf{k}) = \theta$.

We can see that the current matrix $\mathbf{A}_i$ allows for contemporaneous correlations between ν_t and ϵ_{it}. This way of expanding can be easily adapted for many different models at almost no cost. Moreover, if SEM is likelihood based, then MLE of PX-SEM, due to separability of the log-likelihood, can be easily obtained. As a result, we expect improvement in terms of overall algorithmic efficiency.

In the other two applications, we follow this logic of building L model through affine transformation. But we will explore more flexible L model by relaxing constraints in matrix $\mathbf{A}$, such as allowing for correlations across periods. By expanding to a larger L model space, we aim to achieve faster convergence.

5 Discrete Choice Models

The second type of model we will discuss is the random effects discrete choice models with persistent and transitory shocks. The discrete choice models are widely used in empirical works on different topics such as labor supply (Hyslop, 1999), consumer demand (Keane et al., 2013), etc. Distinguishing unobserved heterogeneity from the persistent component is of interest for many reasons, but the nonlinearity and the latent feature complicate the estimation. However, the simulation involved in the SEM or PX-SEM makes the two methods suitable for estimating this type of model. Therefore, take a Probit model as an example, we will discuss its estimation procedure. Specifically, we allow for rich structure by including unobserved time-invariant effects, persistent component, and transitory component. We will later compare the performances of PX-SEM and SEM.

O Model:

$$y_{it} = \mathbb{1}(z_{it} > 0),$$

$$z_{it} = \beta' x_{it} + \mu_i + \nu_{it} + \epsilon_{it},$$

$$\nu_{it} = \rho \nu_{i,t-1} + u_{it}$$

where $\mu_i|x \sim N(0, \sigma_\mu^2)$, $u_{it} \sim N(0, \sigma_u^2)$, $\epsilon_{it} \sim N(0,1)$, $v_{i1} \sim N(0,1)$.[8]

For each individual $i \in 1, ..., N$ at period $t \in 1, ..., T$, we observe a vector of independent variable x_{it} of dimension J and a 0-1 discrete dependent variable y_{it}, whereas z_{it}, individual effect μ_i, persistent component v_{it} are latent variables. Denote the set of unknown parameters by $\theta \equiv (\beta, \sigma_\mu, \rho, \sigma_u)$.

SEM algorithm. For comparison, we first explain the procedure of SEM algorithm. Starting from a guess $\hat{\theta}^{(0)}$, we iterate between the E-step and M-step on $s = 0, 1, 2, ..., S$ until the convergence of $\hat{\theta}^{(s)}$ to the stationary distribution:

1. Stochastic E step: Draw z_i, μ_i, and v_i from posterior distribution $f_O(z_i, \mu_i, v_i|y_i, x_i; \hat{\theta}^{(s)})$
2. M step:
   - $\hat{\beta}^{(s+1)} = (\sum x_{it} x_{it}')^{-1} (\sum x_{it}(z_{it} - \mu_i - v_{it}))$
   - $\hat{\sigma}_\mu^{(s+1)} = \widehat{\text{std}}(\mu_i)$
   - $\hat{\rho}^{(s+1)} = (\sum v_{i,t-1} v_{i,t-1}')^{-1} (\sum v_{i,t-1} v_{it})$
   - $\hat{\sigma}_u^{(s+1)} = \widehat{\text{std}}(v_{it} - \hat{\rho} v_{i,t-1})$

PX-SEM algorithm. To implement PX-SEM algorithm, we need to build a proper L model. Defining $x_i = [x_{i1}' \ ... \ x_{iT}']'$, $v_i = [v_{i1} \ ... \ v_{iT}]'$, $\epsilon_i = [\epsilon_{i1} \ ... \ \epsilon_{iT}]'$, we choose the following L model:

L Model

$$y_{it} = \mathbb{1}(z_{it} > 0),$$

$$z_{it} = \gamma_t' x_i + \mu_i + v_{it} + \epsilon_{it},$$

$$\begin{bmatrix} \mu_i \\ v_i \\ \epsilon_i \end{bmatrix} = \mathbf{pA} \begin{bmatrix} \mu_i^* \\ v_i^* \\ \epsilon_i^* \end{bmatrix} + \mathbf{B} x_i,$$

$$v_{it}^* = \rho v_{i,t-1}^* + u_{it},$$

$$\mu_i^*|x \sim N(0, \sigma_\mu^2), u_{it} \sim N(0, \sigma_u^2), \epsilon_{it}^* \sim N(0,1), v_{i1}^* \sim N(0,1)$$

subject to $\frac{1}{\mathbf{p}} \times (C\mathbf{B} + \gamma) = I_{T\times T} \otimes \beta'$, $C\mathbf{A}\Sigma\mathbf{A}'C' = C\Sigma C'$, and $\mathbf{p} > 0$, where $C = [\vec{1}_{T\times 1} \ I_{T\times T} \ I_{T\times T}]$, $\Sigma = \text{cov}([\mu_i^* \ v_i^{*\prime} \ \epsilon_i^{*\prime}]')$, $\gamma \equiv [\gamma_1 \ ... \ \gamma_T]'$ and $\mathbf{A}$ is a lower triangular matrix with positive diagonal entries. In addition to unknown parameter θ from the O model, L model contains a vector of auxiliary parameters $K \equiv [\text{vech}(\mathbf{A})', \text{vec}(\mathbf{B})', \mathbf{p}]'$.

Our logic of model expansion is very straightforward. We assume that latent variables μ_i^*, v_i^*, and ϵ_i^* follow the same distributions as their counterparts in the O model. However,

[8] Extensions including 1) Logit, that is $\epsilon_{it} \sim Logistic$, and 2) Allowing for dependence of μ_i and v_{i1} on x_{i1}, that is $\mu_i|x \sim N(\beta_\mu x_{i1}, \sigma_\mu^2)$, and $v_{i1}|x \sim N(\beta_v x_{i1}, 1)$ are discussed in Appendix E.

the E-step draws $[\mu_i \ \nu_i' \ \epsilon_i']$ are possibly the result of an affine map acting on the vector $[\mu_i^* \ \nu_i^{*\prime} \ \epsilon_i^{*\prime}]$. In this way, we expand the original model by allowing for linear correlations among μ_i, ν_i, ϵ_i, and x_i.

Therefore, related to condition (1), it is easy to verify that L model nests the O model: when $\mathbf{B} = 0_{(2T+1)\times(J\times T)}$, $\mathbf{A} = I_{(2T+1)\times(2T+1)}$, $\underrightarrow{\mathbf{p}} = 1$, the two models coincide $f_O(y_i, z_i, \mu_i, \nu_i|x_i; \theta) = f_L(y_i, z_i, \mu_i, \nu_i|x_i; \theta, \mathbf{A} = I, \mathbf{B} = \underrightarrow{0}, \mathbf{p} = 1)$.

The L model has two main constraints. On top of identification issue, more importantly, the constraints let us obtain the reduction function easily. The first constraint $\frac{1}{\mathbf{p}} \times (C\mathbf{B} + \gamma) = I_{T\times T} \otimes \beta'$ is on the coefficient x_{it} and makes sure that the conditional mean of z_i on x_i in the L model is simple $\mathbf{p}$ times the conditional mean in the O model. The second constraint $C\mathbf{A}\Sigma\mathbf{A}'C' = C\Sigma C'$ guarantees that conditional covariance of z_i on x_i is the $\mathbf{p}^2$ times the conditional covariance in the O model.[9]

Therefore, related to condition (2), it is easy to verify that there exits the reduction function, which is $R(\theta, K) = \theta$, such that $f_O(y_i|x_i; R(\theta, K)) = f_L(y_i|x_i; \theta, K)$.

Intuitively, expanding the O model through these auxiliary parameters help "constrain" the potential violation of model assumptions by E-step draws under some parameter guesses. For example, matrix $\mathbf{B}$ takes care of the linear correlation between observables x_i and latent draws which could happen when the guess $\hat{\beta}$ in E-step is much smaller than the true value in absolute level. Similarly, matrix $\mathbf{A}$ allows for linear correlation among μ_i, ν_{i1}, u_{it} and ϵ_i; Scalar $\mathbf{p}$ scales up and down z_{it} and allows the var(ϵ_{it}) to be different from 1.

Finally, we discuss the procedures to implement PX-SEM algorithm. We start from a guess $\hat{\theta}^{(0)}$, and iterate between the following E-step and PX-M-step on $s = 0, 1, 2, ..., S$ until the convergence of $\hat{\theta}^{(s)}$ to the stationary distribution:

1. Stochastic E step: Draw z_i, μ_i, and ν_i from posterior distribution $f_O(z_i, \mu_i, \nu_i|y_i, x_i; \hat{\theta}^{(s)})$
2. PX-M step: Estimate L model

$$\hat{\theta}^{(s+1)}, \hat{K} = \arg\min_{\theta, K} \sum_i \Psi(\theta, K; y_i, z_i, x_i, \mu_i, \nu_i)$$

In the following paragraphs, we explain the moments we used to estimate $\hat{\theta}^{(s+1)}$. The detailed specifications are presented in Appendix D.

- $\mathbf{p}\beta$: combining the first constraint, $E(x_{it}(z_{it} - \mathbf{p}\beta' x_{it})) = 0$

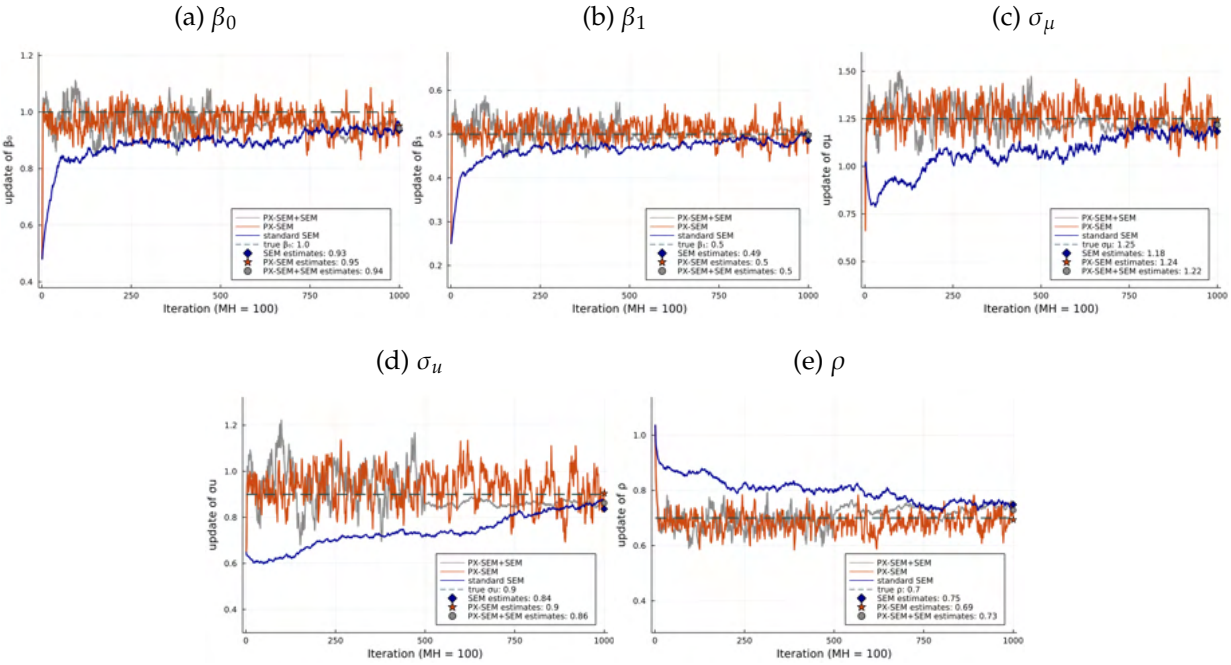
[9] We can rewrite the model as $z_i = \gamma x_i + C[\mu_i \ \nu_i' \ \epsilon_i']' = \mathbf{p}\beta' x_{it} + \mathbf{p}C\mathbf{A}[\mu_i^* \ \nu_i^{*\prime} \ \epsilon_i^{*\prime}]'$.

- **B**: $E(x_i([\mu_i\ \nu_i'\ \epsilon_i'] - x_i'\mathbf{B}')) = 0$
- σ_μ, **p**, Σ, **A**: $C\mathbf{A}\Sigma\mathbf{A}'C' = C\Sigma C'$, and moment constraints on Σ
- ρ, σ_u: $E(\nu^*_{i,t-1}(\nu^*_{it} - \rho\nu^*_{i,t-1})) = 0$, $\text{var}(\nu^*_{it} - \rho\nu^*_{i,t-1}) = \sigma_u^2$

Simulation Results. Now we conduct simulations and compare SEM and PX-SEM algorithms. The true parameters of DGP are: $\beta = [1.0; 0.5]$, $\sigma_\mu = 1.25$, $\rho = 0.7$, and $\sigma_u = 0.9$.

First, we compare SEM and PX-SEM iterations from an informed guess. The initial guess is decided as follows: 1) $\hat{\beta}^{(0)}$ is the Probit regression coefficients of y_{it} on x_{it}, 2) impose $\hat{\sigma}_\mu^{(0)} = 1$, 3) $\hat{\rho}^{(0)}$,$\hat{\sigma}_u^{(0)}$ are computed from the residual of linear regress y_{it} on x_{it}.[10] In both E-step, we use a random-walk Metropolis-Hastings sampler. The acceptance rate is controlled to be between 20% and 40%.

Figure 3: SEM and PX-SEM iterations of $\theta^{(s)}$ from informed guesses



Notes: Complete iterations of SEM (blue solid line), PX-SEM (orange solid line), and PX-SEM + SEM(grey solid line, 500 iterations each) based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond), PX-SEM estimates (orange star), and PX-SEM+SEM (grey circle) are all based on the average of last 250 iterations. Based on informed initial guess.

Figure 3 presents the estimation results of one simulation with $N = 5000$ and $T = 8$. Specifically, we plot the M-step updates $\hat{\theta}^{(s)}$ for 1000 iterations ($S = 1000$). The blue line shows each update of the SEM algorithm, whereas the orange one represents the

[10]Specifically, regress $|err_1| * sign(y_{it} - 0.1) - err_2 * \hat{\sigma}_\mu^{(0)}$ on x_{it} and keep residual $\widehat{res}_{it}$, set $\hat{\rho}^{(0)} = \frac{1}{N(T-1)} \sum \frac{\widehat{res}_{i,t-1}\widehat{res}_{it}}{\widehat{res}^2_{i,t-1}}$, and $\hat{\sigma}_u^{(0)} = \widehat{\text{std}}(\widehat{res}_{it} - \hat{\rho}^{(0)}\widehat{res}_{i,t-1})$

PX-SEM updates. We also combine PX-SEM for 500 iterations with the SEM algorithm with another 500 iterations, and use the grey line to represent the result. The green dash line indicates the true value. We take the average of the last 250 iterations as the final estimates ($S^0 = 250$), which are represented by the blue diamond and orange star for SEM and PX-SEM algorithms, respectively. We can see that starting from the same initial guess, the PX-SEM algorithm converges almost immediately to the region near the true value, whereas SEM moves much slower, especially for $\hat{\sigma}_\mu^{(s)}$, $\hat{\sigma}_u^{(s)}$, and $\hat{\rho}^{(s)}$, and does not converge within 750 iterations.[11] In the case of PX-SEM+SEM, it is obvious that the variation along iterations reduces dramatically once we switch to the SEM algorithm. But since we take the average of the last 250 updates as estimates, there is no significant difference between PX-SEM and PX-SEM+SEM in this example.

Next, we compare the iterations based on random initial guesses. This strategy is common in practice. As it usually is hard to know what are "good" initial guesses and to avoid the method converging to a local maximum, researchers often implement SEM algorithms from many different initial guesses and choose one based on certain criteria, such as likelihood.

In Figure 4, we show that the PX-SEM algorithm still greatly increases the convergence speed: starting from some random initial guesses, despite some jumps at the beginning, the PX-SEM algorithm converges to the region near the true value within 100 iterations, whereas the SEM algorithm, with the same initial guesses, does not seem to converge within 500 iterations.

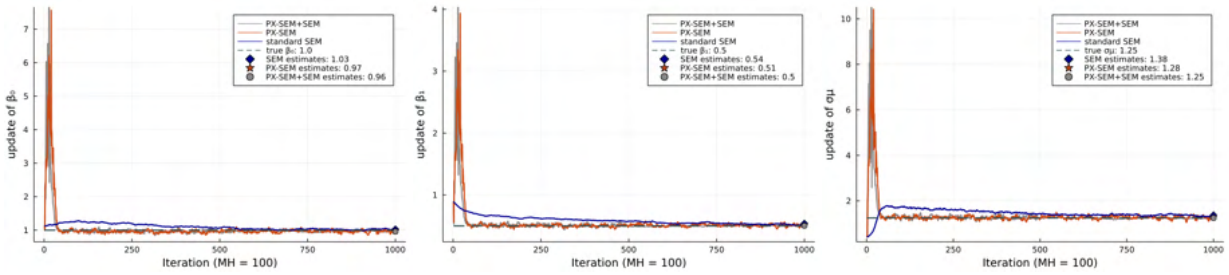
To have a better view of the details, we plot the last 900 iterations in Figure 5. As shown in the figure, SEM seems to converge at the very end of the iterations. As a result, the point estimates taken as the average of the last 500 iterations are relatively far from the true values compared to others. As for PX-SEM+SEM, after switching to the SEM algorithm, the grey line shows much fewer variations which is in line with our discussion on the statistical properties of moments-based PX-SEM.

Considering that this type of exercise will be conducted many times in practice, the save of time could be enormous. More simulation results are presented in Appendix G.

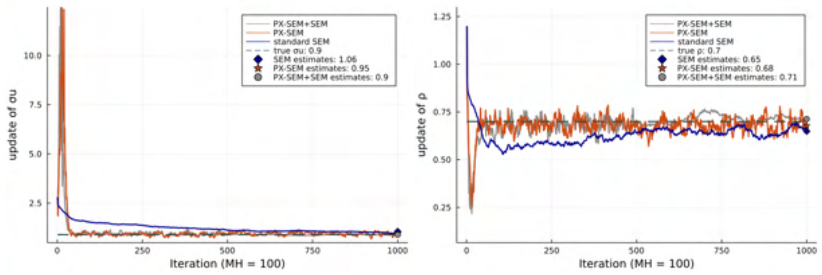
[11] In Appendix F, we show that in some other simulations, it could take more than 2000 iterations for SEM to converge.

Figure 4: SEM and PX-SEM iterations of $\theta^{(s)}$ from random initial guesses

(a) β_0 (b) β_1 (c) σ_μ



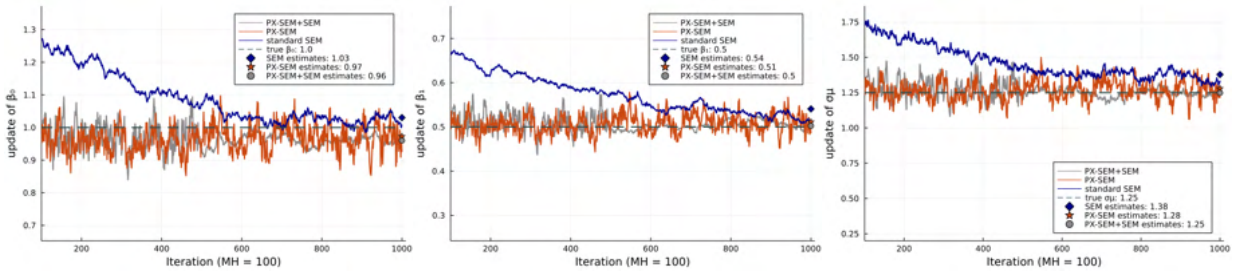
(d) σ_u (e) ρ



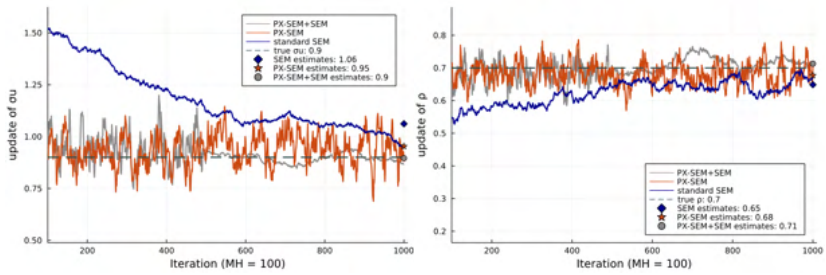
Notes: Complete iterations of SEM (blue solid line), PX-SEM (orange solid line), and PX-SEM + SEM(grey solid line, 500 iterations each) based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond), PX-SEM estimates (orange star), and PX-SEM+SEM (grey circle) are all based on the average of last 500 iterations. Random initial guess from lognormal distribution.

Figure 5: SEM and PX-SEM iterations of $\theta^{(s)}$ from random initial guesses

(a) β_0 (b) β_1 (c) σ_μ



(d) σ_u (e) ρ



Notes: Last 900 iterations of SEM (blue solid line), PX-SEM (orange solid line), and PX-SEM + SEM(grey solid line, 500 iterations each) based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond), PX-SEM estimates (orange star), and PX-SEM+SEM (grey circle) are all based on the average of last 500 iterations. Random initial guess from lognormal distribution. The complete iterations are presented in Figure 4.

6 Quantile Models

The last example we will discuss is the persistent-transitory dynamic quantile processes with individual effects based on Arellano et al., 2017.[12] The model does not impose restrictions on the distributions of individual effects and transitory shock. Moreover, its flexibility in the dynamics of the persistent component allows for features including nonlinear persistence. The model has been applied to topics including income dynamics, firm dynamics, health dynamics, etc.

Denote the τth conditional quantile of v_{it} given $v_{i,t-1}$ as $Q(v_{i,t-1}, \tau)$ for each $\tau \in (0,1)$. The O model to be estimated is as follows:

O Model:

$$y_{it} = \mu_i + v_{it} + \epsilon_{it},$$

$$v_{it} = Q(v_{i,t-1}, u_{it}), (u_{it}|\mu_i, u_{i,t-1}, u_{i,t-2}, ...) \sim \text{Uniform}(0,1), t = 2, ..., T$$

where ϵ_{it} has zero mean, i.i.d. over time, and independent of v_i and μ_i. Individual effect μ_i is assumed to be independent from ϵ_i and v_i.

Unlike the canonical permanent-transitory process, this model allows for a much more flexible conditional distribution of persistent component v_{it} such that it could generate features like nonlinear persistence.

To estimate this model, we follow Arellano et al., 2017 and empirically specify the components as follows:

$$Q(v_{i,t-1}, \tau) = \sum_{k=0}^{K} \gamma_k^Q(\tau)\varphi_k(v_{i,t-1})$$

$$Q_\epsilon(\tau) = \gamma^\epsilon(\tau), \quad Q_{v_1}(\tau) = \gamma^{v_1}(\tau), \quad Q_\mu(\tau) = \gamma^\mu(\tau)$$

where φ_k is Hermite polynomials up to order K.

In Arellano et al., 2017, the model is estimated using a variation of the SEM algorithm where the M-step consists of a sequence of quantile regressions rather than likelihood optimization for computational convenience. For comparison, we first explain their procedures. Denote θ for all unknown parameters including $\gamma_k^Q(\tau)$, $\gamma^\epsilon(\tau)$, $\gamma^{v_1}(\tau)$, and $\gamma^\mu(\tau)$.[13] Starting from initial guesses $\hat{\theta}^{(0)}$, we iterate the E-step and the M-step until convergence to stationary distribution:

[12] In this example, we remove age effect, and assume that the unobserved heterogeneity only affects the level of y_{it}, but does not interact with v_{it}.

[13] Unknown parameters also include tail parameters. Functions $\gamma(\cdot)$ are piecewise-polynomial interpolating splines on a grid $[\tau_1, \tau_2]$, $[\tau_2, \tau_3]$,...,$[\tau_{L-1}, \tau_L]$. And the tails on $(0, \tau_1]$ and $[\tau_L, 1)$ are modeled using parametric model. Check the Appendix B in Arellano et al., 2017 for more details.

1. Stochastic E step: Draw μ_i and ν_i from posterior distribution $f_O(\mu_i, \nu_i | y_i; \hat{\theta}^{(s)})$
2. M step: Update parameters by computing a series of quantile regressions:

$$\hat{\gamma}^Q(\tau) = \arg \min_{\gamma_0^Q, \gamma_1^Q, \ldots, \gamma_K^Q} \sum_{i=1}^{N} \sum_{t=2}^{T} \rho_\tau \left(\nu_{it} - \sum_{k=0}^{K} \gamma_k^Q \varphi_k(\nu_{i,t-1}) \right)$$

$$\hat{\gamma}^\epsilon(\tau) = \arg \min_{\gamma^\epsilon} \sum_{i=1}^{N} \sum_{t=1}^{T} \rho_\tau \left(\epsilon_{it} - \gamma^\epsilon \right)$$

$$\hat{\gamma}^{\nu_1}(\tau) = \arg \min_{\gamma^{\nu_1}} \sum_{i=1}^{N} \rho_\tau \left(\nu_{i1} - \gamma^{\nu_1} \right)$$

$$\hat{\gamma}^\mu(\tau) = \arg \min_{\gamma^\mu} \sum_{i=1}^{N} \rho_\tau \left(\mu_i - \gamma^\mu \right)$$

PX-SEM algorithm. We take the same strategy and expand the O model targeting linear correlations among μ_i, ν_i, and ϵ_i. To achieve this, we propose the following L model:

L Model:

$$y_{it} = \mu_i + \nu_{it} + \epsilon_{it},$$

$$\begin{bmatrix} \mu_i \\ \nu_i \\ \epsilon_i \end{bmatrix} = \mathbf{A} \begin{bmatrix} \mu_i^* \\ \nu_i^* \\ \epsilon_i^* \end{bmatrix}$$

$$\nu_{it}^* = Q(\nu_{i,t-1}^*, u_{it}), (u_{it} | \mu_i^*, u_{i,t-1}, u_{i,t-2}, \ldots) \sim \text{Uniform}(0,1), t = 2, \ldots, T$$

subject to $C\mathbf{A} = C$, where $C = [\overrightarrow{1}_{T\times 1} \; I_{T\times T} \; I_{T\times T}]$. Similarly, we assume that ϵ_{it}^* has zero mean, i.i.d. over time, and independent of ν_i^* and μ_i^*, μ_i^* is independent from ϵ_i^* and ν_i^*. L model contains a vector of auxiliary parameters $K \equiv \text{vec}(\mathbf{A})$.

The logic of model expansion is the same as before: we assume that variables μ_i^*, ν_i^*, and ϵ_i^* follow the same distributions as their counterparts in the O model, but the E-step draws $[\mu_i \; \nu_i' \; \epsilon_i']'$ are possibility the outcome of an affine transformation of $[\mu_i \; \nu_i^{*\prime} \; \epsilon_i^{*\prime}]'$ with coefficient matrix $\mathbf{A}$ to generate linear correlations among μ_i, ν_i, and ϵ_i. Thus, the condition (1) to implement PX-SEM holds, which is that L model nests O model, since we can always take $\mathbf{A} = I$, and the two models coincide $f_O(y_i, \mu_i, \nu_i; \theta) = f_L(y_i, \mu_i, \nu_i; \theta, \mathbf{A} = I)$.

The constraint on matrix $\mathbf{A}$, $C\mathbf{A} = C$, guarantees that the condition (2) holds: there exists a reduction function $R(\theta, K)$, which is simply $\theta = R(\theta, K)$, such that $f_O(y_i; R(\theta, K)) =$

$f_L(y_i;\theta,K)$.[14] This constraint is tighter than the one in the discrete choice case where we just need to target the first two moments due to normality assumption.

To implement PX-SEM algorithm, we start from initial guesses $\hat{\theta}^{(0)}$, and iterate the E-step and the PX-M-step until convergence to stationary distribution:

1. Stochastic E step: Draw μ_i and ν_i from posterior distribution $f_O(\mu_i,\nu_i|y_i;\hat{\theta}^{(s)})$
2. PX-M step: Estimate L model

$$\hat{\theta}^{(s+1)},\hat{K}=\arg\min_{\theta,K}\sum_i \Psi(\theta,K;y_i,\mu_i,\nu_i)$$

In the following paragraphs, we explain the estimation strategy of L model. The detailed steps are listed in Appendix H.

With the complication of matrix $\mathbf{A}$, we can no longer conduct a series of quantile regressions directly using E-step draws as in the SEM. Considering the number of unknown parameters, joint estimation can be very challenging. The strategy we take is to estimate parameters sequentially: we first obtain estimates of auxiliary parameters $\hat{\mathbf{A}}$ using moment restrictions including zero correlations among μ_i^* and ϵ_i^*, and ν_{it}^* following a time-homogeneous first-order Markov process, and then we estimate Θ by the same series of quantile regressions using $\hat{\mathbf{A}}^{-1}[\mu_i\ \nu_i'\ \epsilon_i']'$.

To estimate matrix $\mathbf{A}$, we add extra restrictions on its form and the specification is as follows:

$$\mathbf{A}=\left[\begin{array}{cccccccc}
\mathbf{a}^{\mu} & a_{01}^{\nu} & \cdots & a_{0T}^{\nu} & a_{01}^{\epsilon} & \cdots & a_{0T}^{\epsilon} \\
\hline
1-a^{\mu} & 1-a_{01}^{\nu}-a_{11}^{\nu} & \cdots & -a_{0T}^{\nu}-a_{1T}^{\nu} & 1-a_{01}^{\epsilon}-\mathbf{a}^{\epsilon} & \cdots & -a_{0T}^{\epsilon}-a_{1T}^{\epsilon} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1-a^{\mu} & -a_{01}^{\nu}-a_{T1}^{\nu} & \cdots & 1-a_{0T}^{\nu}-a_{TT}^{\nu} & -a_{01}^{\epsilon} & \cdots & 1-a_{0T}^{\epsilon}-\mathbf{a}^{\epsilon} \\
\hline
0 & a_{11}^{\nu} & \cdots & a_{1T}^{\nu} & \mathbf{a}^{\epsilon} & \cdots & a_{1T}^{\epsilon} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & a_{T1}^{\nu} & \cdots & a_{TT}^{\nu} & 0 & \cdots & \mathbf{a}^{\epsilon}
\end{array}\right]
\begin{array}{l}
\left.\vphantom{\begin{array}{c}a\\a\end{array}}\right\}\mathbf{A1}_{1\times(2T+1)} \\
\left.\vphantom{\begin{array}{c}a\\a\\a\end{array}}\right\}\mathbf{A2}_{T\times(2T+1)} \\
\left.\vphantom{\begin{array}{c}a\\a\\a\end{array}}\right\}\mathbf{A3}_{T\times(2T+1)}
\end{array}$$

In addition to satisfy $C\mathbf{A}=C$, we also assume that entries in the first column of $\mathbf{A3}$ are all zero, and the submatrix made of the last T columns, $\mathbf{A3}_{(:,T+1:2T+1)}$, is an upper triangular matrix with diagonal elements all equal to $\mathbf{a}^{\epsilon}$.

These extra restrictions are made mainly to simplify the estimation. It is shown in the Appendix H that given any value of $\mathbf{a}^{\mu}$ and $\mathbf{a}^{\epsilon}$, we have close form solution for

[14] We can rewrite the model as $y_i=C[\mu_i\ \nu_i'\ \epsilon_i']'=C\mathbf{A}[\mu_i^*\ \nu_i^{*\prime}\ \epsilon_i^{*\prime}]'=C[\mu_i^*\ \nu_i^{*\prime}\ \epsilon_i^{*\prime}]'$.

$\mathbf{A}(\mathbf{a}^\mu, \mathbf{a}^\epsilon)$ and $[\mu_i^* \ \nu_i^{*\prime} \ \epsilon_i^{*\prime}]' = \mathbf{A}^{-1}(\mathbf{a}^\mu, \mathbf{a}^\epsilon)[\mu_i' \ \nu_i' \ \epsilon_i']'$ using moment restrictions including that μ_i and ϵ_i are uncorrelated with all the other elements of the model. We use extra moments including that ν_i follows the first-order markov process to obtain $\hat{\mathbf{a}}^\mu$ and $\hat{\mathbf{a}}^\epsilon$. The estimators, which are the GMM estimators weighted by the Lagrangian multiplier, are not as efficient as the optimal ones, but again we face simpler optimization problem. Once we have $\hat{\mathbf{a}}^\mu$, $\hat{\mathbf{a}}^\epsilon$, and thus $[\hat{\mu}_i^* \ \hat{\nu}_i^{*\prime} \ \hat{\epsilon}_i^{*\prime}]' = \hat{\mathbf{A}}^{-1}(\hat{\mathbf{a}}^\mu, \hat{\mathbf{a}}^\epsilon)[\mu_i' \ \nu_i' \ \epsilon_i']'$, we do a list of quantile regressions to estimate θ. The steps are summarized as follows:

1. Obtain $\hat{A}$
   (a) Given each $\mathbf{a}^\mu$ and $\mathbf{a}^\epsilon$, obtain $\hat{\mathbf{A}}(\mathbf{a}^\mu, \mathbf{a}^\epsilon; \mu_i, \nu_i, \epsilon_i)$ and $\left[\hat{\mu}_i^* \quad \hat{\nu}_i^{*\prime} \quad \hat{\epsilon}_i^{*\prime}\right]' = \hat{\mathbf{A}}^{-1}\left[\mu_i \quad \nu_i' \quad \epsilon_i'\right]'$ using moments $\text{cov}(u_i^*, \nu_{it}^*) = 0$, $\text{cov}(u_i^*, \epsilon_{it}^*) = 0$, $\text{cov}(\epsilon_{ik}^*, \epsilon_{it}^*) = 0$, $\forall t, k \in [1, ..., T]$ and $t \neq k$
2. Compute $\hat{\mathbf{a}}^\mu, \hat{\mathbf{a}}^\epsilon = \arg\min_{a^\mu, a^\epsilon} \sum_i G(\hat{\mu}_i^*, \hat{\nu}_i^*, \hat{\epsilon}_i^*)$
3. Compute $\left[\hat{\mu}_i^* \quad \hat{\nu}_i^{*\prime} \quad \hat{\epsilon}_i^{*\prime}\right]' = \hat{\mathbf{A}}(\hat{\mathbf{a}}^\mu, \hat{\mathbf{a}}^\epsilon; \mu_i, \nu_i, \epsilon_i)^{-1}\left[\mu_i \quad \nu_i' \quad \epsilon_i'\right]'$, and update parameters by a series of quantile regressions

$$\hat{\gamma}_L^Q(\tau) = \arg\min_{\gamma_0^Q, \gamma_1^Q, ..., \gamma_K^Q} \sum_{i=1}^N \sum_{t=2}^T \rho_\tau\left(\hat{\nu}_{it}^* - \sum_{k=0}^K \gamma_k^Q \varphi_k(\hat{\nu}_{i,t-1}^*)\right)$$

$$\hat{\gamma}_L^\epsilon(\tau) = \arg\min_{\gamma^\epsilon} \sum_{i=1}^N \sum_{t=1}^T \rho_\tau\left(\hat{\epsilon}_{it}^* - \gamma^\epsilon\right)$$

$$\hat{\gamma}_L^{\nu_1}(\tau) = \arg\min_{\gamma^{\nu_1}} \sum_{i=1}^N \rho_\tau\left(\hat{\nu}_{i1}^* - \gamma^{\nu_1}\right)$$

$$\hat{\gamma}_L^\mu(\tau) = \arg\min_{\gamma^\mu} \sum_{i=1}^N \rho_\tau\left(\hat{\mu}_i^* - \gamma^\mu\right)$$

where $G(\cdot)$ is a known function that is informative of the distance between the empirical distribution of $\left[\hat{\mu}_i^* \quad \hat{\nu}_i^{*\prime} \quad \hat{\epsilon}_i^{*\prime}\right]'$ and assumptions on $\left[\mu_i^* \quad \nu_i^{*\prime} \quad \epsilon_i^{*\prime}\right]'$, including the distance from $\hat{\nu}_i^*$ to time-homogeneous first-order Markov chain. The details in each step are described in Appendix H. In Appendix I we also discuss an alternative PX-SEM method for the quantile models.

Simulation Results. Some preliminary simulation results are shown in Appendix J. The general conclusion is that PX-SEM could accelerate the movement towards the true value, and the PX-SEM+SEM performs the best in the exercise. There still exist some problems mainly in step 2 such as the numerical stability associated with nonlinear optimization.

7 Conclusion

In this paper, we develop PX-SEM algorithms by combining the parameter-expansion technique with the stochastic EM algorithm. The method consists of an E-step where we draw latent variables from posterior distribution given observables, and an M-step where we expand the original model to a larger model, estimate the larger model, and finally, reduce to the original model space. The intuition is that the model for the latent variables contains extra information that can be exploited to correct the M step in progressing from a coarse parameter guess to a more accurate one. To implement the method, one needs to build a suitable L model for the original model. This paper focuses on nonlinear panel data models and develops PX-SEM algorithms for three types of models: 1) dynamic factor models, 2) discrete choice models with individual effects, persistent and transitory components, and 3) persistent-transitory dynamic quantile processes. The simulation results show that PX-SEM significantly improves the convergence speed.

References

Arcidiacono, Peter and John Bailey Jones (2003) "Finite mixture distributions, sequential likelihood and the EM algorithm," *Econometrica*, 71 (3), 933–946.

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 Stéphane Bonhomme (2016) "Nonlinear panel data estimation via quantile regressions," *The Econometrics Journal*, 19 (3), C61–C94, http://www.jstor.org/stable/45172103.

Bai, Jushan, Serena Ng et al. (2008) "Large dimensional factor analysis," *Foundations and Trends® in Econometrics*, 3 (2), 89–163.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977) "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, 39 (1), 1–22.

Diebolt, Jean and Gilles Celeux (1993) "Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions," *Stochastic Models*, 9 (4), 599–613.

Geweke, John (1977) "The dynamic factor analysis of economic time series," *Latent variables in socio-economic models*.

Hyslop, Dean R (1999) "State dependence, serial correlation and heterogeneity in intertemporal labor force participation of married women," *Econometrica*, 67 (6), 1255–1294.

Keane, Michael P et al. (2013) *Panel data discrete choice models of consumer demand*: Nuffield College.

Lavielle, Marc and Cristian Meza (2007) "A parameter expansion version of the SAEM algorithm," *Statistics and Computing*, 17 (2), 121–130.

Liu, Chuanhai, Donald B Rubin, and Ying Nian Wu (1998) "Parameter expansion to accelerate EM: the PX-EM algorithm," *Biometrika*, 85 (4), 755–770.

Liu, Jun S and Ying Nian Wu (1999) "Parameter expansion for data augmentation," *Journal of the American Statistical Association*, 94 (448), 1264–1274.

Nielsen, Søren Feodor (2000) "The stochastic EM algorithm: estimation and asymptotic results," *Bernoulli*, 457–489.

Stock, James H and Mark W Watson (2006) "Forecasting with many predictors," *Handbook of economic forecasting*, 1, 515–554.

——— (2011) "Dynamic factor models."

Wei, Greg CG and Martin A Tanner (1990) "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms," *Journal of the American statistical Association*, 85 (411), 699–704.

APPENDIX to "Estimating Latent-Variable Panel Data Models Using Parameter-Expanded EM Methods"

A Alternative PX-SEM for Discrete Choice Models

Alternative 1. We start from a guess $\hat{\Theta}^{(0)}$, and iterate between the following E-step and PX-M-step on $s = 0, 1, 2, ..., S$ until the convergence of $\hat{\Theta}^{(s)}$ to the stationary distribution:

1. Stochastic E step: Draw y_i^*, μ_i, and ν_i from posterior distribution $f_O(y_i^*, \mu_i, \nu_i | y_i, x_i; \hat{\Theta}^{(s)})$
2. PX-M step:
   (a) Estimate L model:
      i. $\widehat{\mathbf{p}\beta} = (\sum x_{it} x_{it}')^{-1} (\sum x_{it} y_{it}^*)$
      ii. $\hat{\mathbf{B}}$: $\hat{b}_\mu = (\sum x_i x_i')^{-1} (\sum x_i \mu_i)$, $\hat{b}_\nu' = (\sum x_i x_i')^{-1} (\sum x_i \nu_i')$.
      Then define $\tilde{\mu}_i = \mu_i - \hat{b}_\mu' x_i$, $\tilde{\nu}_i = \nu_i - \hat{b}_\nu' x_i$, $\tilde{\epsilon}_i = y_i^* - \widehat{\mathbf{p}\beta}' x_i - \tilde{\mu}_i - \tilde{\nu}_i$, and $\tilde{y}_i = \tilde{\mu}_i + \tilde{\nu}_i + \tilde{\epsilon}_i$
      iii. $\hat{\sigma}_\mu, \hat{\mathbf{p}}, \hat{\rho}, \hat{\sigma}_u = \arg\min_{\sigma_\mu, p, \rho, \sigma_u} \left\| W^{\frac{1}{2}} \left(\text{vec}(C\widehat{\text{cov}}([\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i])C') - \text{vec}(p^2 C \Sigma(\sigma_\mu, \rho, \sigma_u) C') \right) \right\|$
      iv. $\hat{\beta} = \frac{\widehat{\mathbf{p}\beta}}{\hat{\mathbf{p}}}$
      v. $\hat{\Theta}_L = (\hat{\beta}, \hat{\sigma}_\mu, \hat{\rho}, \hat{\sigma}_u)$
   (b) Reduction: identity mapping due to L model constraints $\hat{\Theta}^{(s+1)} = \hat{\Theta}_L$

Alternative 2. With a different L model, we can implement PX-SEM as follows:

L Model

$$y_{it} = \mathbb{1}(y_{it}^* > 0),$$

$$y_{it}^* = \mathbf{p}\beta' x_{it} + \mu_i + \nu_{it} + \epsilon_{it},$$

$$\nu_{it} = \rho \nu_{i,t-1} + u_{it} - \mathbf{a} u_{i,t-1}$$

where $\mu_i | x \sim N(0, \mathbf{p}^2 \sigma_\mu^2)$, $u_{it} \sim N(0, \mathbf{p}^2 \sigma_u^2)$, $\epsilon_{it} \sim N(0, \mathbf{p}^2)$, $\nu_{i1} \sim N(0, \mathbf{p}^2)$.

There are two auxiliary parameters: scalars **a** and **p**. Parameter **p** works the same as other L model we have discussed before, that is allowing the variance of ϵ_{it} to be

different from 1. Parameter **a** allows for correlation across periods between $v_{it} - \rho_{v_{i,t-1}}$ and $v_{it'} - \rho_{v_{i,t'-1}}$ for $t \neq t'$.

We start from a guess $\hat{\Theta}^{(0)}$, and iterate between the following E-step and PX-M-step on $s = 0, 1, 2, ..., S$ until the convergence of $\hat{\Theta}^{(s)}$ to the stationary distribution:

1. Stochastic E step: Draw y_i^*, μ_i, and v_i from posterior distribution $f_O(y_i^*, \mu_i, v_i | y_i, x_i; \hat{\Theta}^{(s)})$
2. PX-M step:
   (a) Estimate L model:
      i. $\widehat{\mathbf{p}\beta} = (\sum x_{it} x_{it}')^{-1} (\sum x_{it} y_{it}^*)$
      ii. $\widehat{\mathbf{p}\sigma}_\mu = \widehat{\text{std}}(\mu_i)$
      iii. $\hat{\rho} = (\sum v_{i,t-2} v_{i,t-1}')^{-1} (\sum v_{i,t-2} v_{it})$
      iv. $\hat{\mathbf{p}} = \sqrt{\widehat{\text{var}}(y_{it}^* - \mathbf{p}\beta' x_{it} - \mu_i - v_{it}) + |\widehat{\text{cov}}(v_{it} - \hat{\rho} v_{i,t-1}, v_{i,t-1} - \hat{\rho} v_{i,t-2})/\hat{\rho}|}$
      v. $\widehat{\mathbf{p}\sigma_u} = \sqrt{\widehat{\text{var}}(v_{it} - \hat{\rho} v_{i,t-1}) - (1 + \hat{\rho}^2) |\widehat{\text{cov}}(v_{it} - \hat{\rho} v_{i,t-1}, v_{i,t-1} - \hat{\rho} v_{i,t-2})/\hat{\rho}|}$
      vi. $\hat{\beta} = \widehat{\mathbf{p}\beta}/\hat{\mathbf{p}}$, $\hat{\sigma}_\mu = \widehat{\mathbf{p}\sigma}_\mu/\hat{\mathbf{p}}$, $\hat{\sigma}_u = \widehat{\mathbf{p}\sigma_u}/\hat{\mathbf{p}}$
      vii. $\hat{\Theta}_L = (\hat{\beta}, \hat{\sigma}_\mu, \hat{\rho}, \hat{\sigma}_u)$
   (b) Reduction: identity mapping due to L model constraints $\hat{\Theta}^{(s+1)} = \hat{\Theta}_L$

B Comparison of PX-SEM Estimators

In this section, we compare different PX-SEM estimators.

O Model:

$$y_i = y_i^* + \epsilon_i, \quad \begin{pmatrix} y_i^* \\ \epsilon_i \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

SEM

1. E: draw y^*
2. M: $\hat{\sigma} = std(y^*)$

PXSEM, larger model 1

L1 Model:

$$y_i = y_i^* + \epsilon_i$$

where $[y_i^* \ \ \epsilon_i]' \sim N(0, \begin{pmatrix} k & 0 \\ 1-k & 1 \end{pmatrix} \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} k & 0 \\ 1-k & 1 \end{pmatrix}') \Leftrightarrow N(0, \begin{pmatrix} \sigma^2 k^2 & \sigma^2 k(1-k) \\ \sigma^2 k(1-k) & (1-k)^2\sigma^2 + 1 \end{pmatrix})$

Note this model is equivalent to $y_i = \frac{1}{k} y_i^* + \epsilon_i$, where $[y_i^* \ \ \epsilon_i]' \sim N(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix})$

1. PX-SEM-L1

(a) E: same as SEM

(b) M: $\hat{k} = \widehat{\text{var}}(y^*)/\widehat{\text{cov}}(y, y^*)$, $\widehat{\sigma} = \widehat{\text{std}}(y^*)/\hat{k}$ (also corresponds to MLE)

PXSEM, larger model 2

L2 Model:

$$y_i = y_i^* + \epsilon_i$$

where $[y_i^* \ \ \epsilon_i]' \sim N(0, \begin{pmatrix} 1 & 1-k \\ 0 & k \end{pmatrix}\begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1-k \\ 0 & k \end{pmatrix}') \Leftrightarrow \text{N}(0, \begin{pmatrix} \sigma^2 + (1-k)^2 & k(1-k) \\ k(1-k) & k^2 \end{pmatrix})$

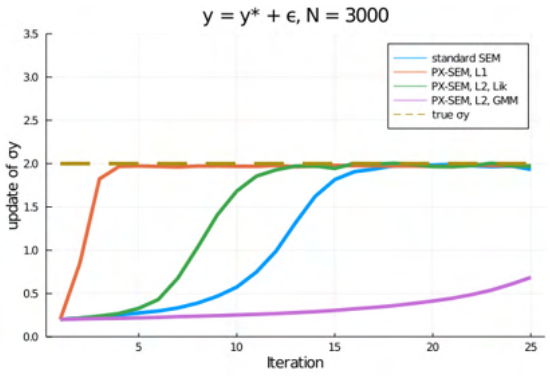
Similar to L1 model, only one auxiliary parameter k. We compare the GMM estimator using variance-covariance matrix with MLE.

1. PX-SEM-L2-1: PX-M step based on GMM estimator
   (a) E: same
   (b) M: $\hat{k} = \widehat{\text{var}}(y - y^*)/\widehat{\text{cov}}(y, y - y^*)$, $\hat{\sigma} = \sqrt{\widehat{\text{var}}(y^*) - (1-\hat{k})^2} \Rightarrow$ if initial guess of σ is very small, then $\hat{\sigma} < \widehat{\text{std}}(y^*)$, the SEM update
2. PX-SEM-L2-2: PX-M step based on MLE
   (a) E: same
   (b) M: MLE

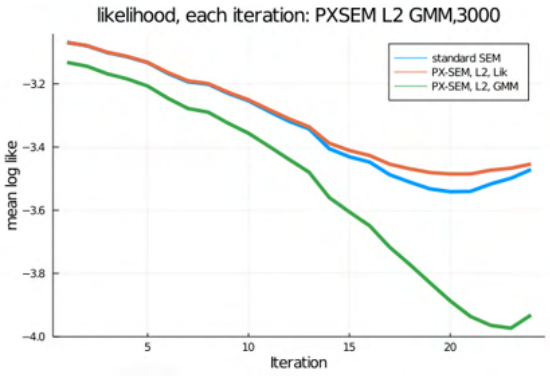
In Figure *B1*, we show both the specification of L models and the PX-M step estimators matter for the performance by comparing SEM, PX-SEM-L1, PX-SEM-L2-1, and PX-SEM-L2-2. The left panel presents the iterations when initial guess of σ is smaller than the true value. First we compare SEM (blue), PX-SEM-L1 (orange), and PX-SEM-L2-2 (green), because in this case both PX-M steps are based on MLE. As expected, PX-SEM algorithms converge faster than SEM. However, even though both L1 and L2 model contain one extra auxiliary parameter, PX-SEM-L1 outperforms PX-SEM-L2-2. This is potentially related to how much flexibility the larger model has to account for the features of the E-step draws.

Figure B1: Comparisons among SEM, PX-SEM-L1, PX-SEM-L2-1, and PX-SEM-L2-2

(a) $\hat{\sigma}^s$ iterations



(b) Log likelihood



Notes: Initial guess smaller than true value. The right panel, the iteration is based on PX-SEM-L2-1. Orange line is $\log f(y, y^*; \hat{\sigma}^{(s+1)}_{L2-2})$*, blue line is* $\log f(y, y^*; \hat{\sigma}^{(s+1)}_{SEM})$*, and green is* $\log f(y, y^*; \hat{\sigma}^{(s+1)}_{L2-1})$*, for* $s = 0, ..., S$

Then we compare SEM (blue), PX-SEM-L2-1 (purple), and PX-SEM-L2-2 (green). As we can see in the figure, PX-SEM-L2-1, with GMM based PX-M-step, performs even worse than SEM. This is not surprising though. It is easy to see why it is happening in this specific case: when the initial guess is smaller, the PX-SEM-L2-1 estimator $\hat{\sigma} = \sqrt{\widehat{\text{var}}(y*) - (1 - \hat{k})^2}$ is smaller than SEM M-step estimator $\widehat{\text{var}}(y*)$. To put it another way, $\hat{\sigma}^{(s)}$ of PX-SEM-L2-1 converges slower. Remember the proof about PX-SEM convergence is based on MLE PX-M step. When we use GMM based estimator in PX-M step, it is likely that, to match some certain moments, the improvement of overall likelihood is worse. This guess is verified in the right panel of Figure B1. Specifically, we implement PX-SEM-L2-1 algorithm. The only difference is that in M-step we also document the complete data likelihood given different estimator:

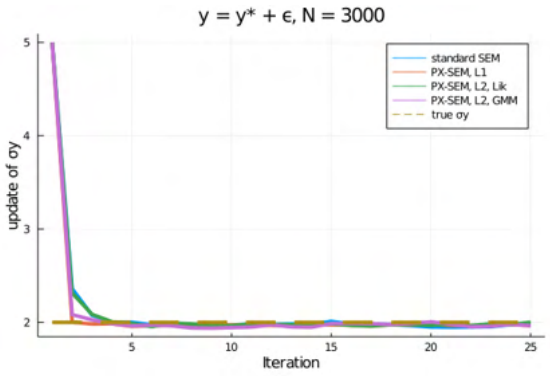
1. E: draw y^* given $\hat{\sigma}^{(s)}_{L2-1}$
2. M: $\hat{k} = \widehat{\text{var}}(y - y^*)/\widehat{\text{cov}}(y, y - y^*)$, $\hat{\sigma}^{(s)}_{L2-1} = \sqrt{\widehat{\text{var}}(y^*) - (1 - \hat{k})^2}$
   1) Compute $\log f(y, y^*; \hat{\sigma}^{(s+1)}_{L2-1})$
   2) Compute $\log f(y, y^*; \hat{\sigma}^{(s+1)}_{SEM})$, where $\hat{\sigma}^{(s+1)}_{SEM} = \widehat{\text{std}}(y^*)$
   3) Compute $\log f(y, y^*; \hat{\sigma}^{(s+1)}_{L2-2})$, where $\hat{\sigma}^{(s+1)}_{L2-2} = \arg\max_\sigma \log f(y, y^*; \sigma)$

As we expected PX-SEM-L2-1 estimator performs worse than SEM in terms of likelihood increment.

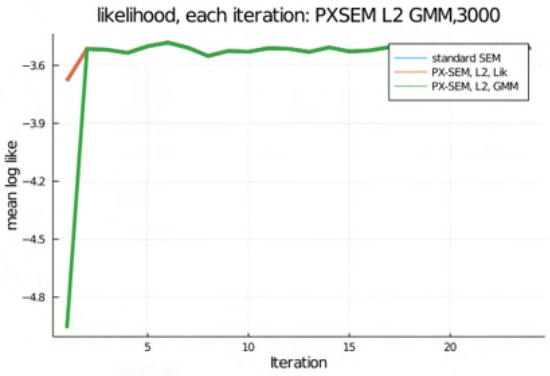
Then we do similar exercise but for an initial guess larger than true value. Figure B2 presents the results. All PX-SEM algorithms perform not worse than SEM algorithm. Specifically, PX-SEM-L2-1, the GMM based one performs even better than PX-SEM-L2-2.

Figure B2: Comparisons among SEM, PX-SEM-L1, PX-SEM-L2-1, and PX-SEM-L2-2

(a) $\hat{\sigma}^s$ iterations



(b) Log likelihood



Notes: Initial guess larger than true value. The right panel, the iteration is based on PX-SEM-L2-1. Orange line is $\log f(y, y^*; \hat{\sigma}^{(s+1)}_{L2-2})$*, blue line is* $\log f(y, y^*; \hat{\sigma}^{(s+1)}_{SEM})$*, and green is* $\log f(y, y^*; \hat{\sigma}^{(s+1)}_{L2-1})$*, for* $s = 0, ..., S$

PXSEM, larger model 3

L3 Model:

$$y_i = y_i^* + \epsilon_i$$

where $[y_i^* \;\; \epsilon_i]' \sim N(0, \begin{pmatrix} k_1 & k_2 \\ 1-k_1 & 1-k_2 \end{pmatrix} \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} k_1 & k_2 \\ 1-k_1 & 1-k_2 \end{pmatrix}')$

This model contains two auxiliary parameters k_1 and k_2. Instead of estimating all parameters jointly, we experiment something closer to the strategy used in Section 6 for quantile models: write down k_2 as a function of k_1 using moment restriction that $cov(y_i^*, \epsilon_i) = 0$, and we have $\hat{\epsilon} = (y^* - k_1 y) * (pinv(y^* - k_1 y) * y)$, $\hat{k}_2 = pinv(\hat{\epsilon}) y^*$. Then we estimate k_1 using different criteria described below:

1. PX-SEM-L3-1: Moments based

$$\hat{k}_1 = \arg\min_{k_1} (\text{var}(\hat{\epsilon}) - 1)^2$$

$$\hat{\sigma} = \widehat{\text{std}}\big(y - (y^* - k_1 y) * (pinv(y^* - \hat{k}_1 y) * y)\big)$$

2. PX-SEM-L3-2: Likelihood based

$$\hat{k}_1 = \max_{k_1} L(y, y^*; k_1, \hat{k}_2(k_1), \hat{\sigma}(k_1))$$

where

$$\hat{\sigma}(k_1) = \widehat{\text{std}}\big(y - (y^* - k_1 y) * (pinv(y^* - k_1 y) * y)\big)$$

Finally

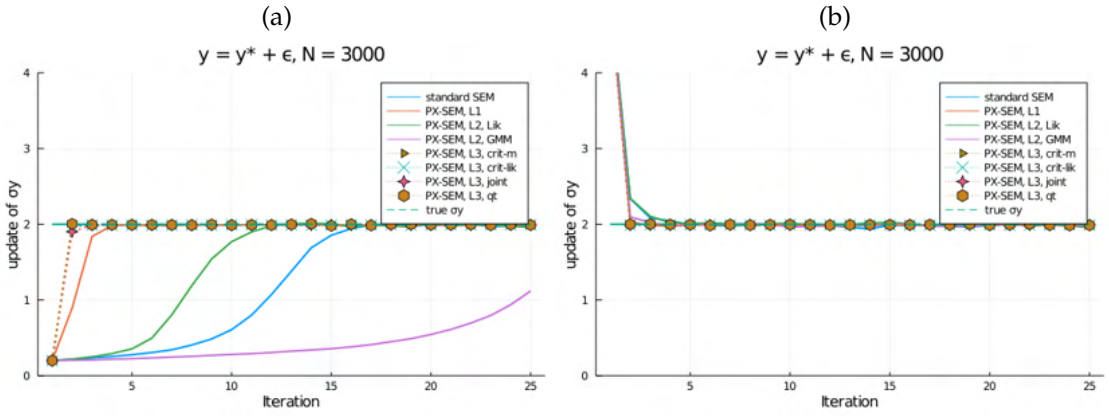
$$\hat{\sigma} = \widehat{\text{std}}\big(y - (y^* - k_1 y) * (pinv(y^* - \hat{k}_1 y) * y)\big)$$

3. PX-SEM-L3-3: Likelihood based

$$\hat{k}_1, \hat{\sigma} = \max_{k_1,\sigma} L(y, y^*; k_1, \hat{k}_2(k_1), \sigma)$$

Finally, we also experiment methods used for quantile models. Specifically, we do not use the information of Normal distribution. The likelihood is quantile based. In PX-M-step, $\hat{k}_1 = \arg\min_{k_1}(\text{var}(\hat{\epsilon}) - 1)^2$, then we update unknown parameters including different quantiles simply look at different quantiles of $y - (y^* - k_1 y) * (pinv(y^* - \hat{k}_1 y) * y)$.

Figure B3: Comparisons among SEM, PX-SEM-L1, PX-SEM-L2*, PX-SEM-L3*, and quantile based one



Notes: The left panel shows results when the initial guesses are smaller than true values. The right panel shows results when the initial guesses are larger than true values.

Comparison among all PX-SEM algorithms is shown in Figure B3. We can see that when the L model is more flexible (L3), at least in this case, even though the M-step estimates are not MLE, the performance is still good, and better than PX-SEM-L1. So there might be a trade-off between the flexibility of L model and PX-M step estimator.

C Asymptotic Properties

In this section, we will derive the dynamics of the PX-SEM updates across iterations. Moreover, we will prove the computational efficiency of the PX-SEM algorithm and show that it provides a consistent estimator without harming statistical efficiency compared to the SEM algorithm. **Currently, the proof applies only to the MLE M-step. However, we anticipate including a discussion on moment-based M-step in a future version of this draft.**

C.1 Dynamics of PX-SEM Updates

Setup. Let $\{Y_i, X_i, Y_i^*\}$ for $i = 1 : N$ be i.i.d. random variables from the O Model distribution $f_O(Y_i|X_i;\theta) = \int_{Y_i^*} f_O(Y_i, Y_i^*|X_i;\theta)dY_i^*$, where $W_i \equiv [Y_i; X_i']'$ is the observable set, Y_i^* is the latent-variable set, and θ is the unknown parameter set to be estimated.The true value, $\overline{\theta}$, satisfies the equation $E(\Psi_O(Y_i^*, W_i;\overline{\theta})) = 0$, where $\Psi_O(\cdot)$ represents the score function of the complete O model. We can easily show that the true value $\overline{\theta}$ also satisfies the equation

$$E\Big(\int \Psi_O(Y_i^*, W_i;\overline{\theta}) f_O(Y_i^*|W_i;\overline{\theta})dY_i^*\Big) = 0 \tag{C1}$$

Define $\hat{\theta}$ as the ML estimator, which is also the solution of the integrated moment restrictions: $\sum_{i=1}^N \int \Psi_O(Y_i^*, W_i;\hat{\theta}) f_O(Y_i^*|W_i;\hat{\theta})dY_i = 0$.

Denote the expanded model, L Model, by $f_L(Y_i|X_i;\theta,K) = \int_{Y_i^*} f_L(Y_i, Y_i^*|X_i;\theta,K)dY_i^*$, where K represents for all auxiliary parameters. The expanded L model needs to satisfy two conditions: (1) L model nests O model: $\exists\ K_0$ such that $f_O(Y_i, Y_i^*|X_i;\theta) = f_L(Y_i, Y_i^*|X_i;\theta,K_0)$, $\forall\theta$, and (2) There is a mapping, the reduction function, from the L Model space to O Model space $\theta = R(\theta_L, K)$ such that the observed data likelihood is preserved $f_O(Y_i|X_i;R(\theta_L,K)) = f_L(Y_i|X_i;\theta_L,K)$, $\forall\theta_L, K$.[1]

Function $\Psi_L^\theta(\cdot)$ represents the score of the L model with respect to θ. Under condition (1), we have $\Psi_L^\theta(Y_i^*, W_i;\theta,K_0) = \Psi_O(Y_i^*, W_i;\theta)$, and thus $E(\Psi_L^\theta(Y_i^*, W_i;\overline{\theta},K_0)) = 0$. Additionally, assumes that K is identified using pseudo-complete data and $\Psi_L^K(\cdot)$ is the score function relative to K, that is $E(\Psi_L^K(Y_i^*, W_i;\overline{\theta},K_0)) = 0$, then we have $E(\Psi_L(Y_i^*, W_i;\overline{\theta},K_0)) = 0$, where $\Psi_L(\cdot) = [\Psi_L^\theta(\cdot); \Psi_L^K(\cdot)]$. Equivalently, we have

$$E\Big(\int \Psi_L(Y_i^*, W_i;\overline{\theta},K_0) f_O(Y_i^*|X_i, Y_i; R(\overline{\theta},K_0))dY_i^*\Big) = 0 \tag{C2}$$

Since in subsection 3.2 we have proved that the observed data likelihood increases in each iteration of PX-SEM, we know the ML estimator $\hat{\theta}$ will also satisfy the following integrated moment restrictions: $\sum_{i=1}^N \int \Psi_L(Y_i^*, W_i;\hat{\theta},K_0) f_O(Y_i^*|W_i;R(\hat{\theta},K_0))dY_i = \sum_{i=1}^N \int \Psi_L(Y_i^*, W_i;\hat{\theta},K_0) f_O(Y_i^*|W_i;\hat{\theta})dY_i = 0$

The general steps are as follows: starting with a guess of unknown parameter $\hat{\theta}^{(0)}$, we iterate the following two steps on $s = 0, 1, 2, ..., S$ until the convergence of $\hat{\theta}^{(s)}$ to the stationary distribution:

1. Stochastic E step: Draw Y_i^* from posterior distribution $f_O(Y_i^*|W_i;\hat{\theta}^{(s)})$

[1] We know reduction function should satisfy $R(\theta, K_0) = \theta$.

2. PX-M step: Update parameters by

   (a) Estimate L model: $\sum_i \Psi_L(Y_i^*, W_i; \hat{\theta}_L^{(s+1)}, \hat{K}^{(s+1)}) = 0$

   (b) Reduction: $\hat{\theta}^{(s+1)} = R(\hat{\theta}_L^{(s+1)}, \hat{K}^{(s+1)})$ subject to $f_O(Y_i|X_i; \hat{\theta}^{(s+1)}) = f_L(Y_i|X_i; \hat{\theta}_L^{(s+1)}, \hat{K}^{(s+1)})$

Following Liu et al., 1998, we use the parameterisation $\Theta = (\theta, K) = (R(\hat{\theta}_L, K), K)$ so that θ represent the PX-SEM update by the end of each iteration. Define the set of true value as $\overline{\Theta} = [R(\overline{\theta}, K_0); K_0] = [\overline{\theta}; K_0]$. Additionally, we have proved that $\hat{\Theta} \equiv [\hat{\theta}; K_0]$ is a fixed point of the PX-SEM iterations where $\hat{\theta}$ is the MLE.

Next, we reparameterize the score function $\Phi_L(\cdot)$ and obtain $\widetilde{\Phi}_L(Y_i^*, W_i, \hat{\theta}^{(s+1)}, \hat{K}^{(s+1)}) = \Phi_L(Y_i^*, W_i, \hat{\theta}_L^{(s+1)}, \hat{K}^{(s+1)})$ given $\hat{\theta}^{(s+1)} = R(\hat{\theta}_L^{(s+1)}, \hat{K}^{(s+1)})$. Then, under the guess $\hat{\Theta}^{(s)} = [\hat{\theta}^{(s)}, K_0]$, PX-SEM finds the update following:

1. Stochastic E step: Draw $Y_i^{*(s)}$ from posterior distribution $f_L(Y_i^{*(s)}|W_i; \hat{\Theta}^{(s)})$
2. PX-M step: Update parameters by solving $\hat{\Theta}^{(s+1)}$ from

$$\sum_{i=1}^{N} \widetilde{\Psi}_L(Y_i^{*(s)}, W_i; \hat{\Theta}^{(s+1)}) = 0$$

Note $\hat{\Theta}^{(s+1)}$ can be seen as MLE of the reparameterized L model $\widetilde{f}_L(Y_i, Y_i^*|X_i, \theta, K) = f_L(Y_i, Y_i^*|X_i, R^{-1}(\theta, K), K)$ where we assume the inverse of the reduction function given K exists.[2]

Following Arellano and Bonhomme, 2016, we rewrite the latent draws using conditional quantile representation, that is

$$Y_i^{*(s)} = Q_{Y^*|W}(W_i, u_i^{(s)}; \hat{\theta}^{(s)}) = Q^L_{Y^*|W}(W_i, u_i^{(s)}; \hat{\theta}^{(s)}, K_0)$$

where $u_i^{(s)}$ is a vector of standard independent uniform draws of same dimension as $Y_i^{*(s)}$, $Q_{Y^*|W}(\cdot)$ is the quantile representation based on O model, and $Q^L_{Y^*|W}(\cdot)$ is the quantile representation based on L model. Then we have the dynamics of PX-SEM updates as:

$$\sum_{i=1}^{N} \widetilde{\Psi}_L(Q^L_{Y^*|W}(W_i, u_i^{(s)}; \hat{\Theta}^{(s)}), W_i; \hat{\Theta}^{(s+1)}) = 0$$

Expand around $\hat{\Theta} \equiv [\hat{\theta}; K_0]$, using the fact that $\hat{\theta}$ goes to $\overline{\theta}$ as N goes to infinity, we have

$$A(\hat{\Theta}^{(s+1)} - \hat{\Theta}) + B(\hat{\Theta}^{(s)} - \hat{\Theta}) + \epsilon^{(s)} = o_p(N^{-(1/2)}) \tag{C3}$$

[2]Score functions: $\sum_{i=1}^N \frac{\partial \ln \widetilde{f}_L(Y_i, Y_i^*|X_i, \hat{\Theta}^{(s+1)})}{\partial \theta} = \sum_{i=1}^N \Psi_L^\theta(Y_i^*, W_i; R^{-1}(\hat{\theta}^{(s+1)}, K^{(s+1)}), K^{(s+1)}) \frac{\partial R^{-1}(\hat{\Theta}^{(s+1)})}{\partial \theta} = \sum_{i=1}^N \widetilde{\Psi}_L^\theta(Y_i^*, W_i; \hat{\Theta}^{(s+1)}) \frac{\partial R^{-1}(\hat{\Theta}^{(s+1)})}{\partial \theta} = 0$; similarly $\sum_{i=1}^N \widetilde{\Psi}_L^\theta(Y_i^*, W_i; \hat{\Theta}^{(s+1)}) \frac{\partial R^{-1}(\hat{\Theta}^{(s+1)})}{\partial K} + \widetilde{\Psi}_L^K(Y_i^*, W_i; \hat{\Theta}^{(s+1)}) = 0$. As long as that $\frac{\partial R^{-1}(\cdot)}{\partial \theta}$ is not zero, we have $\sum_{i=1}^N \widetilde{\Psi}_L^\theta(Y_i^*, W_i; \hat{\Theta}^{(s+1)}) = 0$ and $\sum_{i=1}^N \widetilde{\Psi}_L^K(Y_i^*, W_i; \hat{\Theta}^{(s+1)}) = 0$. Finally we have $\sum_{i=1}^N \widetilde{\Psi}_L(Y_i^*, W_i; \hat{\Theta}^{(s+1)}) = 0$.

where

$$A \equiv \frac{\partial}{\partial \Theta'}\bigg|_{\overline{\Theta}} E(\widetilde{\Psi}_L(Q^L_{Y^*|W}(W_i, u_i; \overline{\Theta}), W_i; \Theta)) = \frac{\partial}{\partial \Theta'}\bigg|_{\overline{\Theta}} E(\widetilde{\Psi}_L(Y_i^*, W_i; \Theta))$$

$$B \equiv \frac{\partial}{\partial \Theta'}\bigg|_{\overline{\Theta}} E(\widetilde{\Psi}_L(Q^L_{Y^*|W}(W_i, u_i; \Theta), W_i; \overline{\Theta})) = \frac{\partial}{\partial \Theta'}\bigg|_{\overline{\Theta}} E\bigg(\int \widetilde{\Psi}_L(Y_i^*, W_i; \overline{\Theta}) f_L(Y_i^*|W_i; \Theta) dY_i^*\bigg)$$

$$\epsilon^{(s)} \equiv \frac{1}{N}\sum_{i=1}^{N} \widetilde{\Psi}_L(Q^L_{Y^*|W}(W_i, u_i; \hat{\Theta}), W_i; \hat{\Theta})$$

We also can prove that

$$A + B = \frac{\partial}{\partial \Theta'}\bigg|_{\overline{\Theta}} E\bigg(\int \widetilde{\Psi}_L(Y_i^*, W_i; \Theta) f_L(Y_i^*|W_i; \Theta) dY_i^*\bigg)$$

C.2 Consistency and Efficiency

Now we can characterize the asymptotic distribution of $\hat{\theta}^{(s)}$ of PX-SEM algorithm. Because $\hat{K}^{(s)} = K_0$, we have

$$\sqrt{N}(\hat{\theta}^{(s)} - \hat{\theta}) = \sum_{l=0}^{\infty} (-A^{-1}B)^l_{[1:J,1:J]} (-A^{-1})_{[1:J,:]} \sqrt{N}\epsilon^{(s-1-l)} + o_p(1)$$

where $(-A^{-1}B)_{[1:J,1:J]}$ represents the submatrix that consists for the first J rows and first J columns of matrix $(-A^{-1}B)$, and $(-A^{-1})_{[1:J,:]}$ is the submatrix made of the first J rows of matrix $-A^{-1}$. Identification requires $-(A)^{-1}B_{[1:J,1:J]} < I$, so that the Gaussian AR(1) limit $\sqrt{N}(\hat{\Theta}^{(s)} - \hat{\Theta})$ conditional on W is stable.

Given that $\epsilon^{(s)} \xrightarrow{d} \mathcal{N}(\vec{0}, \Sigma_\epsilon)$, where

$$\Sigma_\epsilon = E(\widetilde{\Psi}_L(Y_i^*, W_i; \overline{\Theta})\widetilde{\Psi}_L(Y_i^*, W_i; \overline{\Theta})')$$

conditional on W_i, we have

$$\sqrt{N}(\hat{\theta}^{(s)} - \hat{\theta}) \xrightarrow{d} \mathcal{N}(0, \Sigma_1)$$

where

$$\Sigma_1 = \sum_{l=0}^{\infty} (-A^{-1}B)^l_{[1:J,1:J]} (-A^{-1})_{[1:J,:]} \Sigma_\epsilon ((-A^{-1})_{[1:J,:]})' ((-A^{-1}B)^l_{[1:J,1:J]})'$$

Unconditionally, we have

$$\sqrt{N}(\hat{\theta}^{(s)} - \overline{\theta}) = \sqrt{N}(\hat{\theta}^{(s)} - \hat{\theta}) + \sqrt{N}(\hat{\theta} - \overline{\theta}) \xrightarrow{d} \mathcal{N}(0, \Sigma_1 + \Sigma_2^{-1}\Sigma_3\Sigma_2^{-1\prime})$$

where

$$\Sigma_2 = \frac{\partial}{\partial \theta'}\bigg|_{\overline{\theta}} E\bigg(\int \Psi_O(Y_i^*, W_i; \theta) f_O(Y_i^*|W_i; \theta) dY_i^*\bigg)$$

$$\Sigma_3 = E\bigg(\Big(\int \Psi_O(Y_i^*, W_i; \theta) f_O(Y_i^*|W_i; \theta) dY_i^*\Big)\Big(\int \Psi_O(Y_i^*, W_i; \theta) f_O(Y_i^*|W_i; \theta) dY_i^*\Big)'\bigg)$$

C.3 Computational Efficiency

We rewrite the equation (C3) as follows

$$(\hat{\Theta}^{(s+1)} - \hat{\Theta}) = -A^{-1}B(\hat{\Theta}^{(s)} - \hat{\Theta}) - A^{-1}\epsilon^{(s)} + o_p(N^{-(1/2)}) \tag{C4}$$

Following Liu et al., 1998 and Nielsen, 2000, we discuss the parameter $-A^{-1}B$ and variance of innovation $A^{-1}\,\mathrm{var}(\epsilon^{(s)})A^{-1'}$ in further details. First, define $V \equiv A + B$, then we have

$$-A^{-1}B = -A^{-1}(A + B - A) = I - A^{-1}V$$

where $-A$ by definition is the complete-data information matrix, and $-V$ is the observed-data information matrix. It is easy to show that

$$A = \begin{bmatrix} A_{\theta\theta} & A_{\theta K} \\ A_{K\theta} & A_{KK} \end{bmatrix}, V = \begin{bmatrix} V_{\theta\theta} & 0 \\ 0 & 0 \end{bmatrix}$$

where $-A_{\theta\theta}$ and $-V_{\theta\theta}$ are the complete-data information matrix and observed-information matrix based on O model, respectively.

$$\begin{aligned} A_{\theta\theta} &= \frac{\partial}{\partial\theta'}\bigg|_{\overline{\theta}} E\left(\frac{\log \widetilde{f}_L(Y_i^*, W_i; \theta, K_0)}{\partial\theta}\right) = \frac{\partial}{\partial\theta'}\bigg|_{\overline{\theta}} E(\Psi_O(Y_i^*, W_i; \theta)) \\ V_{\theta\theta} &= \frac{\partial}{\partial\theta'}\bigg|_{\overline{\theta}} E\left(\int \frac{\log \widetilde{f}_L(Y_i^*, W_i; \theta, K_0)}{\partial\theta} \widetilde{f}_L(Y_i^*|W_i; \theta, K_0) dY_i^*\right) \\ &= \frac{\partial}{\partial\theta'}\bigg|_{\overline{\theta}} E\left(\int \Psi_O(Y_i^*, W_i; \theta) f_O(Y_i^*|W_i; \theta) dY_i^*\right) \end{aligned}$$

As a benchmark, the SEM algorithm based on O model generates the following dynamics

$$(\hat{\theta}_{SEM}^{(s+1)} - \hat{\theta}) = \big(I - (-A_{\theta\theta}^{-1})(-V_{\theta\theta})\big)(\hat{\theta}_{SEM}^{(s)} - \hat{\theta}) - A_{\theta\theta}^{-1}\epsilon_\theta^{(s)} + o_p(N^{-(1/2)})$$

where $\epsilon_\theta^{(s)} = \epsilon_{[1:J]}^{(s)}$.

Define

$$H = \begin{bmatrix} H_{\theta\theta} & H_{\theta K} \\ H_{K\theta} & H_{KK} \end{bmatrix} \equiv \begin{bmatrix} A_{\theta\theta} & A_{\theta K} \\ A_{K\theta} & A_{KK} \end{bmatrix}^{-1}$$

so that $H_{\theta\theta} = A_{\theta\theta}^{-1} + H_{\theta K}H_{KK}^{-1}H_{K\theta}$, then we have

$$I - A^{-1}V = \begin{bmatrix} I - (-H_{\theta\theta})(-V_{\theta\theta}) & 0 \\ -(-H_{K\theta})(-V_{\theta\theta}) & I \end{bmatrix}$$

and

$$(\hat{\theta}^{(s+1)} - \hat{\theta}) = \big(I - (-H_{\theta\theta})(-V_{\theta\theta})\big)(\hat{\theta}^{(s)} - \hat{\theta}) - A_{[1:J,:]}^{-1}\epsilon^{(s)} + o_p(N^{-(1/2)})$$

Since $-H_{\theta\theta} \geq -A_{\theta\theta}^{-1}$ in semipositive definite order, then the smallest eigenvalue of $(-H_{\theta\theta})(-V_{\theta\theta})$ is at least as large as the smallest eigenvalue of $(-A_{\theta\theta}^{-1})(-V_{\theta\theta})$, we expect that PXSEM on average converge faster than SEM.

D Discrete Choice Model Estimation

In this section, we present in detail the procedures to estimate the L model of the discrete choice model in Section 5.

1. $\widehat{\mathbf{p}\beta} = (\sum x_{it}x_{it}')^{-1}(\sum x_{it}z_{it})$
2. $\hat{\mathbf{B}}$: $\hat{b}_\mu = (\sum x_i x_i')^{-1}(\sum x_i \mu_i)$, $\hat{b}_\nu' = (\sum x_i x_i')^{-1}(\sum x_i \nu_i')$.
   Then define $\tilde{\mu}_i = \mu_i - \hat{b}_\mu' x_i$, $\tilde{\nu}_i = \nu_i - \hat{b}_\nu' x_i$, $\tilde{\epsilon}_i = z_i - \widehat{\mathbf{p}\beta}' x_i - \tilde{\mu}_i - \tilde{\nu}_i$, and $\tilde{z}_i = \tilde{\mu}_i + \tilde{\nu}_i + \tilde{\epsilon}_i$
3. $\hat{\sigma}_\mu, \hat{\mathbf{p}} = \arg\min_{\sigma_\mu, p} \left\| W^{\frac{1}{2}} \left(\mathrm{vec}(\hat{\Sigma}_\nu) - \mathrm{vec}(\Sigma_\nu(\rho(\sigma_\mu, p), \sigma_u(\sigma_\mu, p))) \right) \right\|$,
   where $\hat{\Sigma}_\nu = \widehat{\mathrm{var}}(\tilde{z}_i) - \sigma_\mu^2 p^2 \overrightarrow{1}_{T\times 1} \overrightarrow{1}_{1\times T} - p^2 I_{T\times T}$,
   $\hat{\nu}_i^* = \mathrm{chol}(\hat{\Sigma})\, \mathrm{chol}(\widehat{\mathrm{cov}}([\tilde{\mu}_i\ \tilde{\nu}_i'\ \tilde{\epsilon}_i']'))^{-1}[\tilde{\mu}_i\ \tilde{\nu}_i'\ \tilde{\epsilon}_i']'$,
   $\rho(\sigma_\mu, p) = (\sum \hat{\nu}_{i,t-1}^* \hat{\nu}_{i,t-1}^{*\prime})^{-1}(\sum \hat{\nu}_{i,t-1}^* \hat{\nu}_{i,t}^*)$,
   $\sigma_u(\sigma_\mu, p) = \widehat{\mathrm{std}}(\hat{\nu}_{it}^* - \rho \hat{\nu}_{i,t-1}^*)$
4. $\hat{\rho} = (\sum \hat{\nu}_{i,t-1}^* \hat{\nu}_{i,t-1}^{*\prime})^{-1}(\sum \hat{\nu}_{i,t-1}^* \hat{\nu}_{i,t}^*)$, $\hat{\sigma}_u = \widehat{\mathrm{std}}(\hat{\nu}_{it}^* - \hat{\rho}\hat{\nu}_{i,t-1}^*)$, $\hat{\beta} = \frac{\widehat{\mathbf{p}\beta}}{\hat{\mathbf{p}}}$,
   where $[\hat{\mu}_i^*\ \hat{\nu}_i^{*\prime}\ \hat{\epsilon}_i^{*\prime}]' = \mathrm{chol}(\hat{\Sigma}(\hat{\sigma}_\mu, \hat{\mathbf{p}}))\, \mathrm{chol}(\widehat{\mathrm{var}}([\tilde{\mu}_i\ \tilde{\nu}_i'\ \tilde{\epsilon}_i']'))^{-1}[\tilde{\mu}_i\ \tilde{\nu}_i'\ \tilde{\epsilon}_i']'$
5. $\hat{\theta}^{(s+1)} = (\hat{\beta}, \hat{\sigma}_\mu, \hat{\rho}, \hat{\sigma}_u)$

where Σ is the variance-covariance matrix of $[\mu_i^*\ \nu_i^{*\prime}\ \epsilon_i^{*\prime}]'$, $\Sigma = \mathrm{var}([\mu_i^*\ \nu_i^{*\prime}\ \epsilon_i^{*\prime}]')$, and Σ_ν is the variance-covariance matrix of ν_i^*, $\Sigma_\nu = \mathrm{var}(\nu_i^*)$

Specifically, the estimator $\hat{\sigma}_\mu$ and $\hat{\mathbf{p}}$ in step 3 are obtained through comparing the following two objects:

- From the model, we have

  $$\mathrm{var}\left(C\left([\mu_i\ \nu_i'\ \epsilon_i']' - \mathbf{B}x_i\right)\right) = \mathrm{var}\left(C\left(\mathbf{p}\mathbf{A}[\mu_i^*\ \nu_i^{*\prime}\ \epsilon_i^{*\prime}]'\right)\right)$$

  which implies

  $$C\mathrm{var}([\mu_i\ \nu_i'\ \epsilon_i']' - \mathbf{B}x_i)C' = \mathbf{p}^2 C\mathbf{A}\Sigma\mathbf{A}'C'$$

  Moreover, with the model constraint $C\mathbf{A}\Sigma\mathbf{A}'C' = C\Sigma C'$, we have

  $$C\mathrm{var}([\mu_i\ \nu_i'\ \epsilon_i']' - \mathbf{B}x_i)C' = \mathbf{p}^2 C\Sigma C'$$

  We can write down Σ_ν as a function of $\mathbf{p}$ and σ_μ

  $$\Sigma_\nu = C\mathrm{var}([\mu_i\ \nu_i'\ \epsilon_i']' - \mathbf{B}x_i)C' - \sigma_\mu^2 \mathbf{p}^2 \overrightarrow{1}_{T\times 1} \overrightarrow{1}_{1\times T} - \mathbf{p}^2 I_{T\times T}$$

  Therefore, Σ can be writen as a function of $\mathbf{p}$ and σ_μ

  $$\Sigma = \begin{bmatrix} \mathbf{p}^2\sigma_\mu^* & 0 & 0 \\ 0 & \Sigma_\nu & 0 \\ 0 & 0 & \mathbf{p}^2 I \end{bmatrix}$$

- From the model, we have

$$\text{var}\left([\mu_i\ \nu_i'\ \epsilon_i']' - \mathbf{B}x_i\right) = \text{var}\left(\mathbf{pA}[\mu_i^*\ \nu_i^{*\prime}\ \epsilon_i^{*\prime}]'\right)$$

  which implies

$$\text{var}\left([\mu_i\ \nu_i'\ \epsilon_i']' - \mathbf{B}x_i\right) = \mathbf{p}^2\mathbf{A}\Sigma\mathbf{A}'$$

  With the constraint that matrix $\mathbf{A}$ being a lower triangular matrix with positive diagonal entries, we have

$$\mathbf{pA} = \text{chol}(\text{var}([\mu_i\ \nu_i'\ \epsilon_i']' - \mathbf{B}x_i))\,\text{chol}(\Sigma)^{-1}$$

  and $(\mathbf{pA})^{-1}_{[2:T+1,:]}([\mu_i\ \nu_i'\ \epsilon_i']' - \mathbf{B}x_i)$ follow AR(1) process.

  We could write down Σ_ν as a function of ρ and σ_u using AR(1) structure, where ρ and σ_u satisfy the moment restrictions that $(\mathbf{pA})^{-1}_{[2:T+1,:]}([\mu_i\ \nu_i'\ \epsilon_i']' - \mathbf{B}x_i)$ follow AR(1) process.

Estimators $\hat{\sigma}_\mu$ and $\hat{\mathbf{p}}$ are chosen to minimize the distance between the two matrix.[3,4]

E Discrete Choice Model Extensions

In this section, we will briefly discuss two extensions of the discrete choice model of the main text. The first one is the Probit regression allowing for the dependence of individual effect and initial persistent component to depend on regressors of initial period. The second case is the Logit regression with the same latent-variable structure.

Probit with dependence. The model to be estimated is as follows:

O Model:

$$y_{it} = \mathbb{1}(y_{it}^* > 0),$$

$$y_{it}^* = \beta' x_{it} + \mu_i + \nu_{it} + \epsilon_{it},$$

$$\nu_{it} = \rho\nu_{i,t-1} + u_{it}$$

where $\mu_i|x \sim N(\beta_\mu x_{i1}, \sigma_\mu^2)$, $u_{it} \sim N(0, \sigma_u^2)$, $\epsilon_{it} \sim N(0,1)$, $\nu_{i1}|x \sim N(\beta_\nu x_{i1}, 1)$.

[3]Different estimators of L model can be exploited as well. For example joint optimization: $\hat{\sigma}_\mu, \hat{\mathbf{p}}, \hat{\rho}, \hat{\sigma}_u = \arg\min_{\sigma_\mu,p,\rho,\sigma_u} \left\| W^{\frac{1}{2}} \left(\text{vec}(C\widehat{\text{cov}}([\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i])C') - \text{vec}(p^2 C\Sigma(\sigma_\mu, \rho, \sigma_u)C') \right) \right\|$. In Appendix A we discuss alternative PX-SEM algorithms for discrete models.

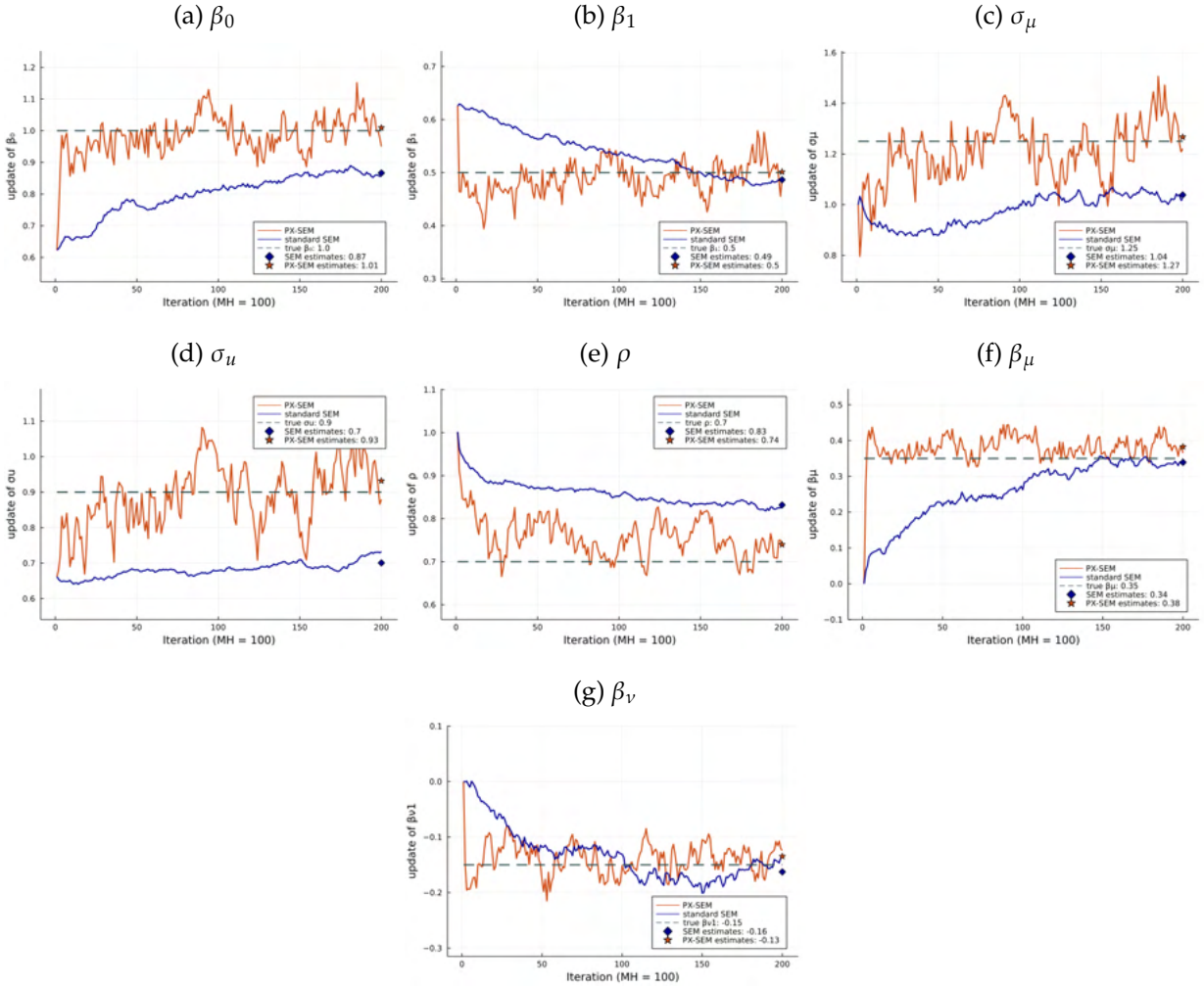
[4]M-step estimator could affect the convergence performance. In some cases, usually associated with some certain initial guesses, the PX-SEM with a GMM estimator in M-step could perform worse than SEM estimator. The reason is that the increment of complete data likelihood based on the GMM estimator might not outperform the increment of the SEM case. But we would expect with a more flexible L model, this becomes less likely to happen. In Appendix B, we explain in detail this problem with the toy model.

For each individual $i \in 1, ..., N$ at period $t \in 1, ..., T$, we observe a $K \times 1$ dimension vector x_i and a 0-1 discrete dependent variable y_i, whereas y_i^*, individual effect μ_i, persistent component v_i are latent variables. Denote the set of unknown parameters by $\Theta = (\beta, \sigma_\mu, \rho, \sigma_u, \beta_\mu, \beta_v)$.

Proposed L model and its estimation will be added in a later draft

We show results:

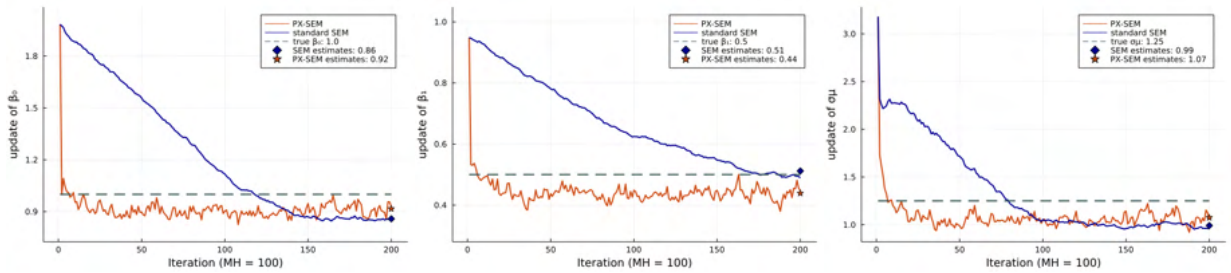
Figure E1: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses



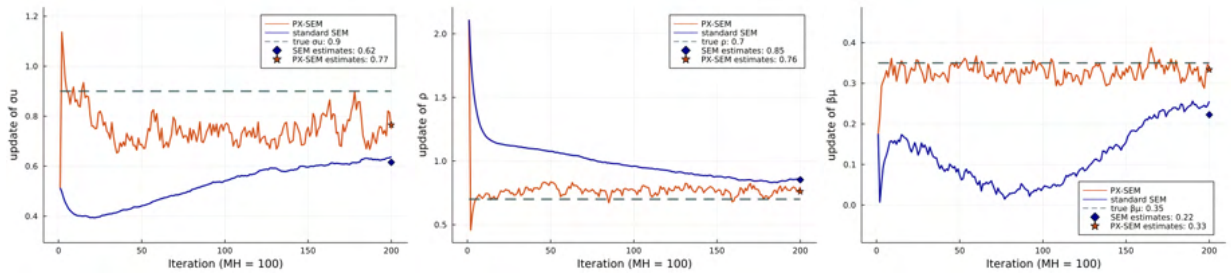
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 50 iterations. Based on informed initial guess.

Figure E2: SEM and PX-SEM iterations of $\Theta^{(s)}$ from random guesses

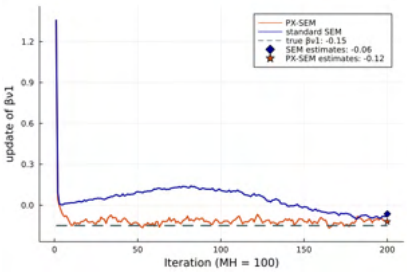
(a) β_0 (b) β_1 (c) σ_μ



(d) σ_u (e) ρ (f) β_μ



(g) β_ν



Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 50 iterations. Based on informed initial guess.

Logit. The model to be estimated is as follows:

O Model:

$$y_{it} = \mathbb{1}(y_{it}^* > 0),$$

$$y_{it}^* = \beta' x_{it} + \mu_i + \nu_{it} + \epsilon_{it},$$

$$\nu_{it} = \rho \nu_{i,t-1} + u_{it}$$

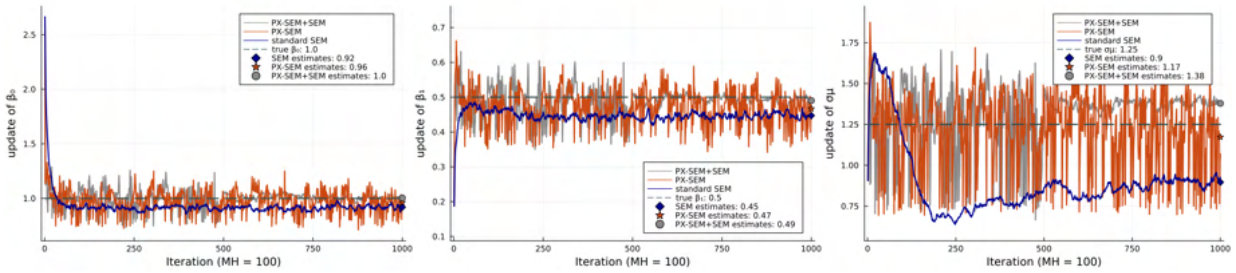
where $\mu_i|x \sim N(0, \sigma_\mu^2)$, $u_{it} \sim N(0, \sigma_u^2)$, $\epsilon_{it} \sim Logistic$, $\nu_{i1}|x \sim N(0,1)$.

For each individual $i \in 1, ..., N$ at period $t \in 1, ..., T$, we observe a $K \times 1$ dimension vector x_i and a 0-1 discrete dependent variable y_i, whereas y_i^*, individual effect μ_i, persistent component ν_i are latent variables. Denote the set of unknown parameters by $\Theta = (\beta, \sigma_\mu, \rho, \sigma_u)$.

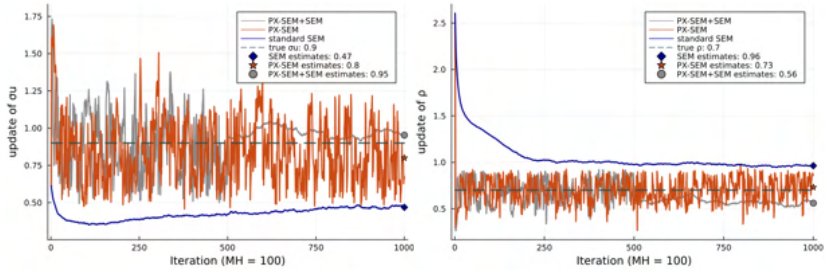
Proposed L model and its estimation will be added in a later draft

Figure E3: SEM and PX-SEM iterations of $\Theta^{(s)}$ from random guesses

(a) β_0 (b) β_1 (c) σ_μ



(d) σ_u (e) ρ



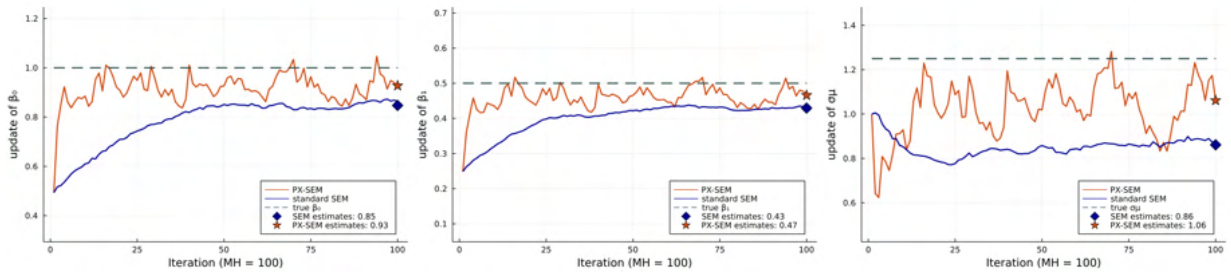
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 250 iterations. Based on informed initial guess.

F Discrete Models with More Iterations

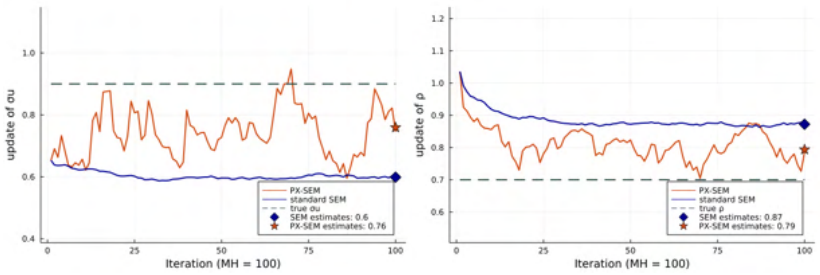
First I show only the 100 iterations. The result is very interesting. If we only conduct SEM algorithm for 100 iterations, it seems the changes along the iterations are rather small since 50th iteration such that one might thought it already converges. However, as we know the true value, we understand it is not the case. The results show that in this simulation, SEM does not converge within 2000 iterations.

Figure F1: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses

(a) β_0 (b) β_1 (c) σ_μ



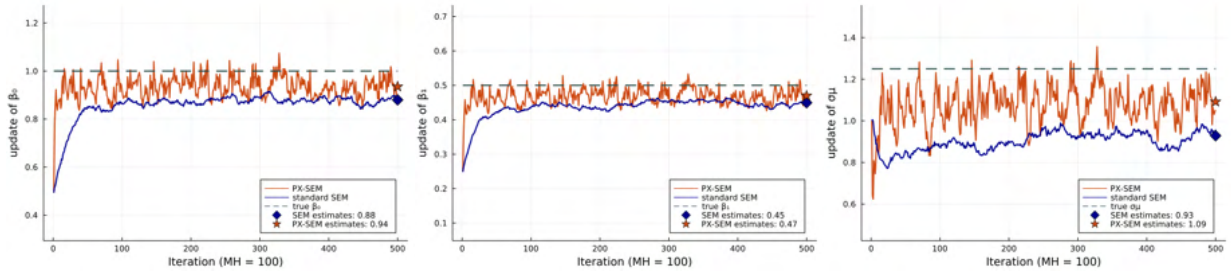
(d) σ_u (e) ρ



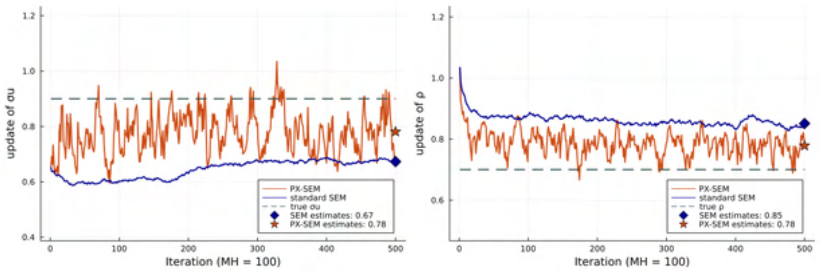
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 50 iterations. Based on informed initial guess.

Figure F2: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses, zoom to $[0, 500]$

(a) β_0 (b) β_1 (c) σ_μ



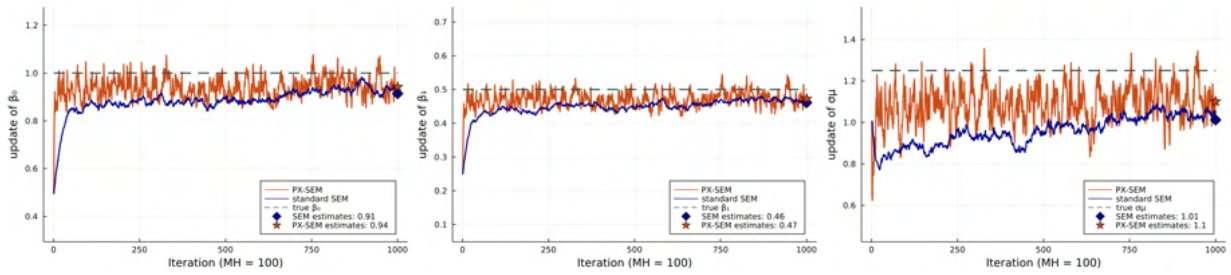
(d) σ_u (e) ρ



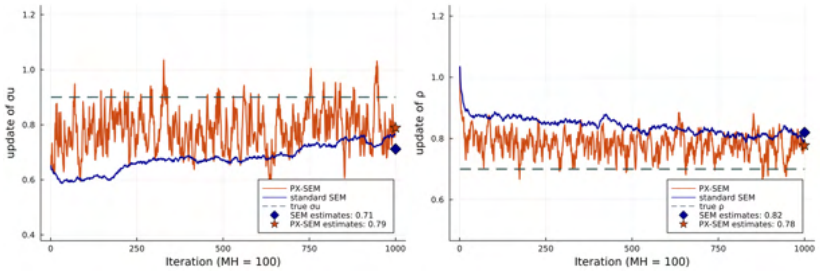
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 250 iterations. Based on informed initial guess.

Figure F3: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses, zoom to $[0, 1000]$

(a) β_0 (b) β_1 (c) σ_μ



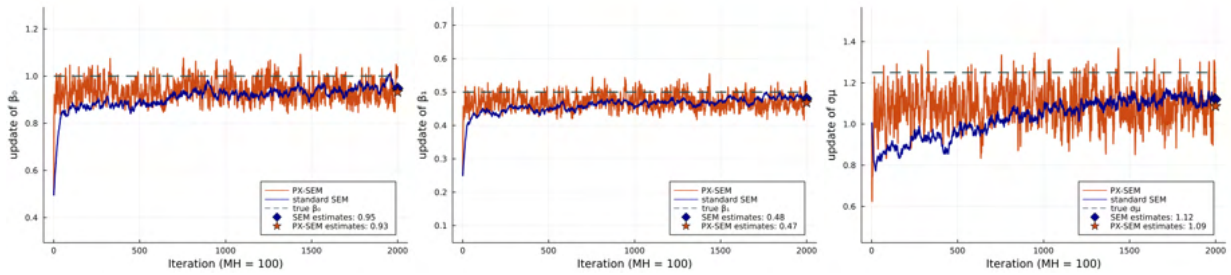
(d) σ_u (e) ρ



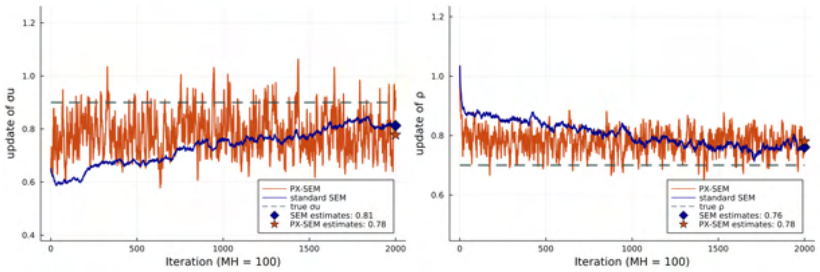
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 500 iterations. Based on informed initial guess.

Figure F4: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses, zoom to $[0, 2000]$

(a) β_0 (b) β_1 (c) σ_μ



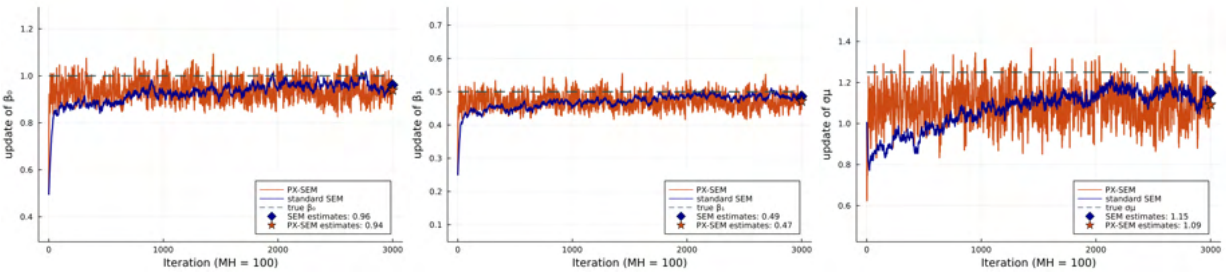
(d) σ_u (e) ρ



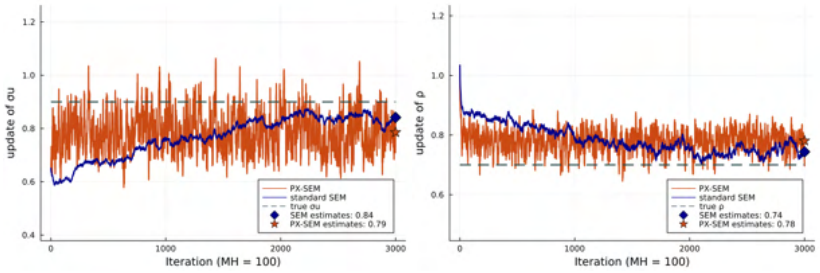
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 500 iterations. Based on informed initial guess.

Figure F5: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses, zoom to $[0, 3000]$

(a) β_0 (b) β_1 (c) σ_μ



(d) σ_u (e) ρ



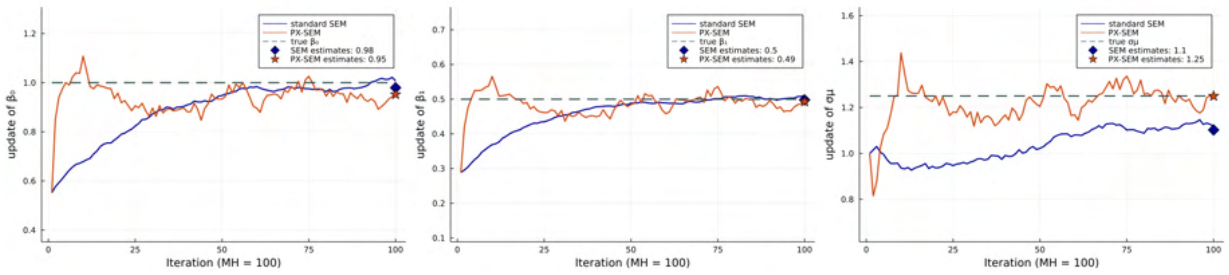
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 1000 iterations. Based on informed initial guess.

G More Simulations of Discrete Models

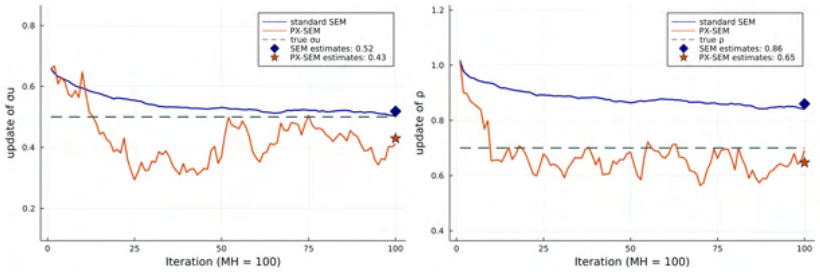
Informed Initial Guesses

Figure G1: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses

(a) β_0 (b) β_1 (c) σ_μ



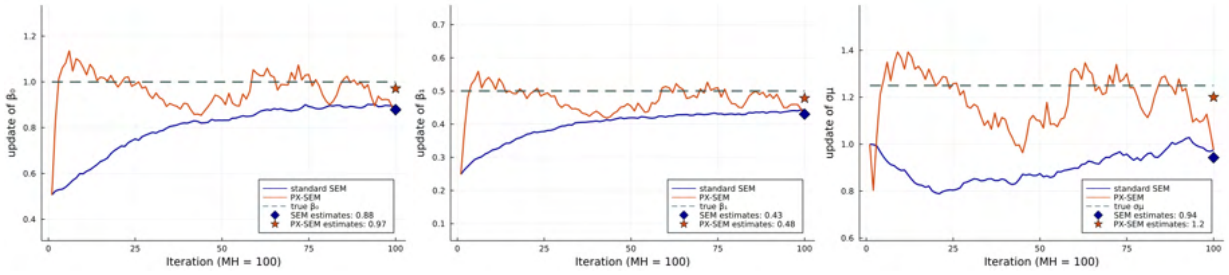
(d) σ_u (e) ρ



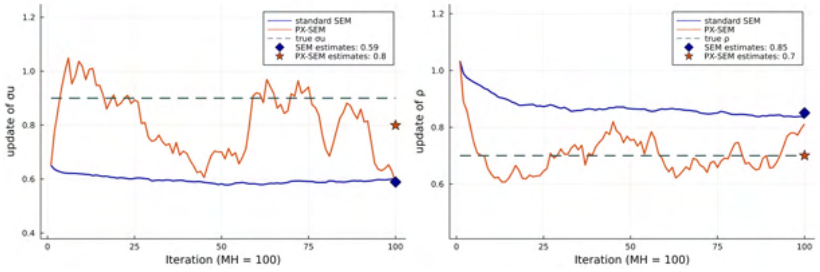
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 50 iterations. Based on informed initial guess.

Figure G2: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses

(a) β_0 (b) β_1 (c) σ_μ



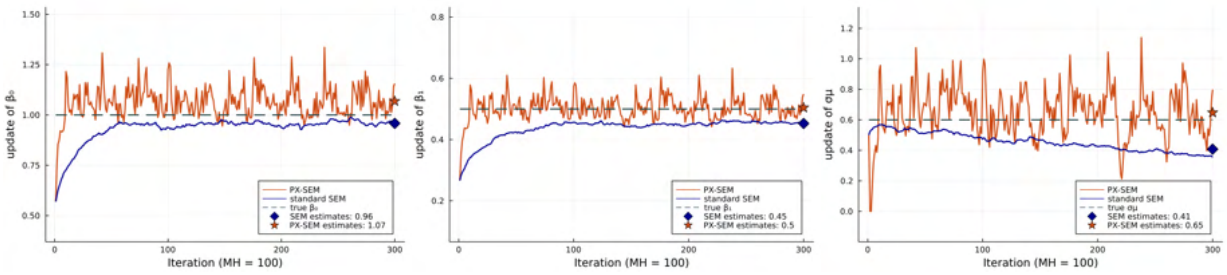
(d) σ_u (e) ρ



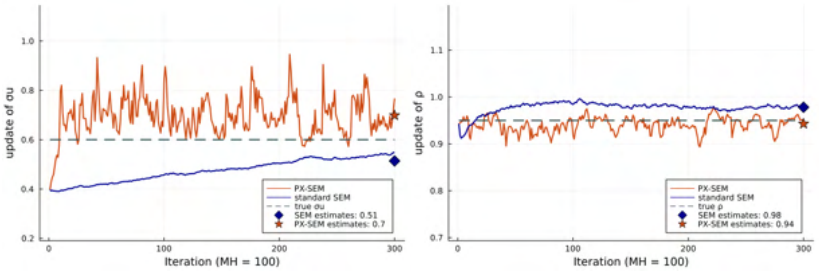
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 50 iterations. Based on informed initial guess.

Figure G3: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses

(a) β_0 (b) β_1 (c) σ_μ



(d) σ_u (e) ρ

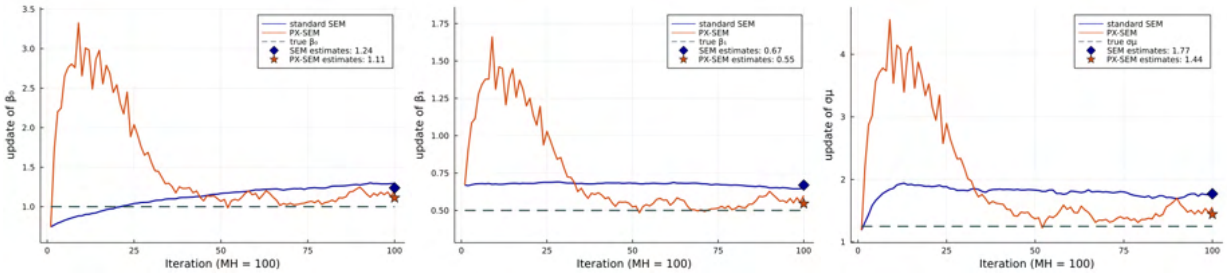


Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 300 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 150 iterations. Based on informed initial guess.

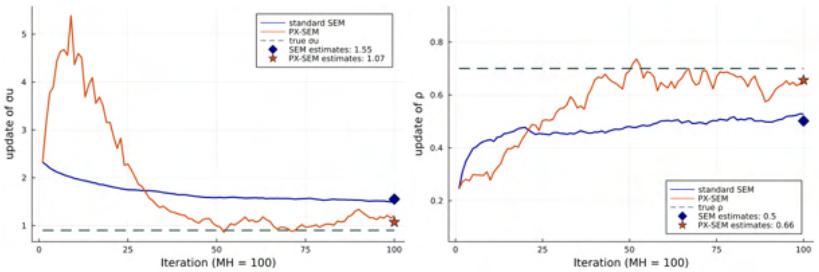
Random Initial Guesses

Figure G4: SEM and PX-SEM iterations of $\Theta^{(s)}$ from random initial guesses

(a) β_0 (b) β_1 (c) σ_μ



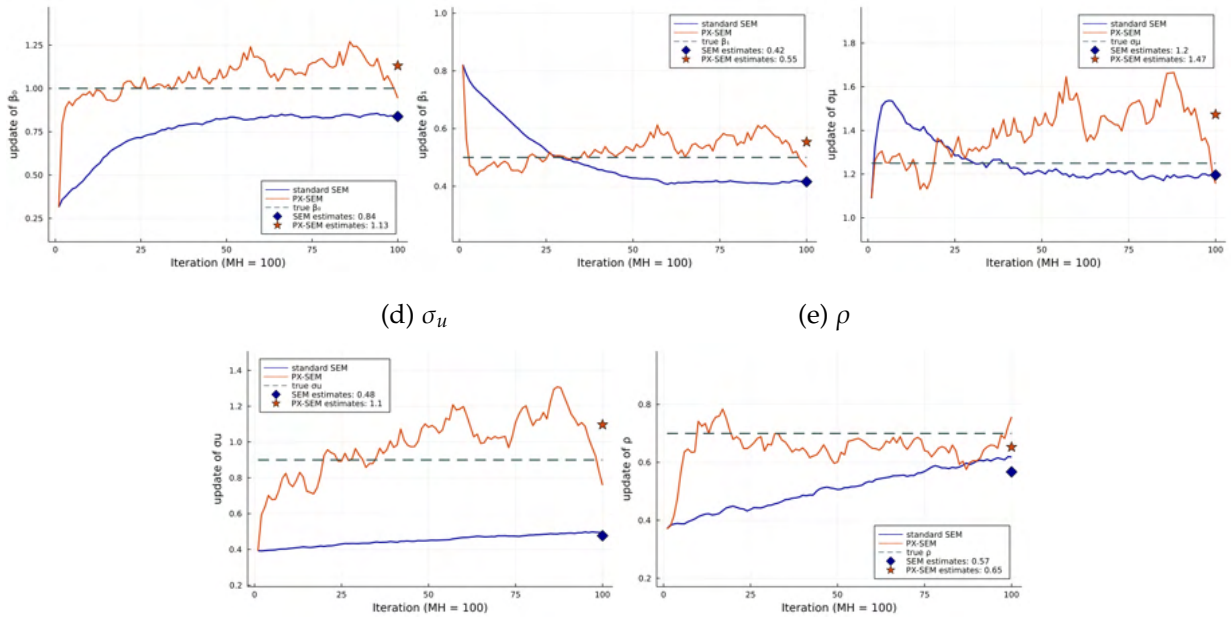
(d) σ_u (e) ρ



Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 50 iterations. Random initial guess from lognormal distribution

Figure G5: SEM and PX-SEM iterations of $\Theta^{(s)}$ from random initial guesses

(a) β_0 (b) β_1 (c) σ_μ

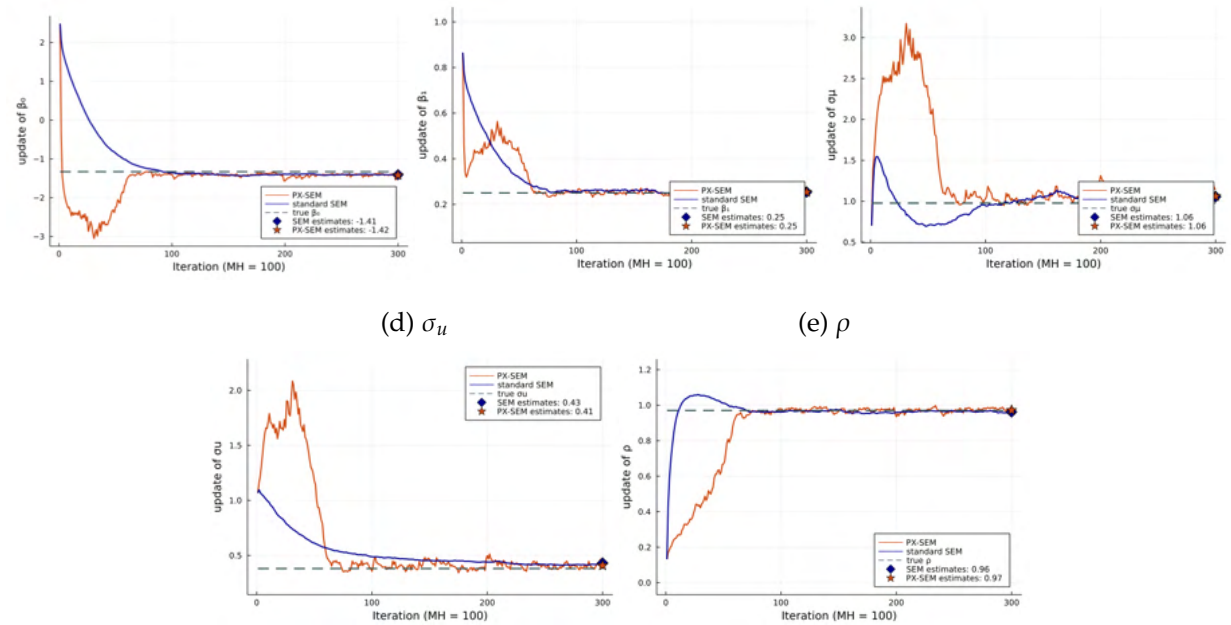


(d) σ_u (e) ρ

Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 50 iterations. Random initial guess from lognormal distribution

Figure G6: SEM and PX-SEM iterations of $\Theta^{(s)}$ from random initial guesses

(a) β_0 (b) β_1 (c) σ_μ

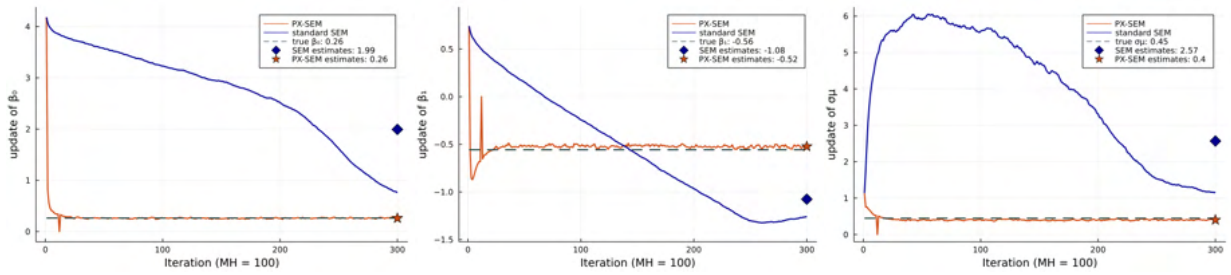


(d) σ_u (e) ρ

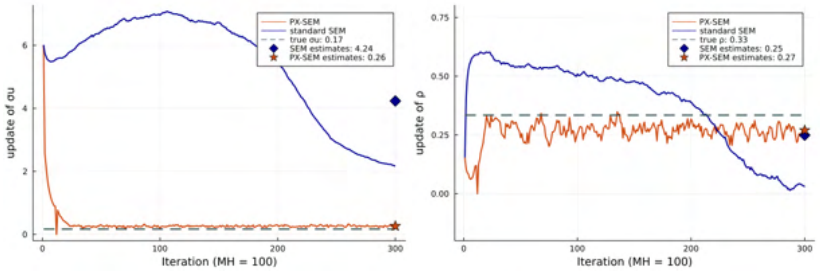
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 300 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 150 iterations. Random initial guess from lognormal distribution

Figure G7: SEM and PX-SEM iterations of $\Theta^{(s)}$ from random initial guesses

(a) β_0 (b) β_1 (c) σ_μ



(d) σ_u (e) ρ



Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 300 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 150 iterations. Random initial guess from lognormal distribution

H PX-M Step of Quantile Models Estimation

In this section, I will first present the moment conditions used to estimate the matrix **A** of L model and then discuss in the detail the M-step procedures.

The moment conditions we use can be devided into three categories. The first category include moments mostly related to matrix **A**; the second and the third categories include moments related to dynamics of ν^*_{it}.

Matrix A. With the specification that

$$\mathbf{A} = \left[\begin{array}{cccccccc} \mathbf{a}^\mu & a^\nu_{01} & \cdots & a^\nu_{0T} & a^\epsilon_{01} & \cdots & a^\epsilon_{0T} \\ \hline 1-a^\mu & 1-a^\nu_{01}-a^\nu_{11} & \cdots & -a^\nu_{0T}-a^\nu_{1T} & 1-a^\epsilon_{01}-\mathbf{a}^\epsilon & \cdots & -a^\epsilon_{0T}-a^\epsilon_{1T} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1-a^\mu & -a^\nu_{01}-a^\nu_{T1} & \cdots & 1-a^\nu_{0T}-a^\nu_{TT} & -a^\epsilon_{01} & \cdots & 1-a^\epsilon_{0T}-\mathbf{a}^\epsilon \\ \hline 0 & a^\nu_{11} & \cdots & a^\nu_{1T} & \mathbf{a}^\epsilon & \cdots & a^\epsilon_{1T} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & a^\nu_{T1} & \cdots & a^\nu_{TT} & 0 & \cdots & \mathbf{a}^\epsilon \end{array}\right] \begin{array}{l} \} \mathbf{A1}_{1\times(2T+1)} \\ \} \mathbf{A2}_{T\times(2T+1)} \\ \} \mathbf{A3}_{T\times(2T+1)} \end{array}$$

and define $\delta_{it}^{\mu} \equiv \mu_i - \mathbf{a}^{\mu} y_{it}$, we can easily solve for $[\frac{1}{T}\sum_t y_{it}\ \delta_i^{\mu\prime}\ \epsilon_i']'$ and have

$$\begin{bmatrix} \frac{1}{T}\sum_t y_{it} \\ \delta_i^{\mu} \\ \epsilon_i \end{bmatrix} = \mathbf{B} \begin{bmatrix} \mu_i^* \\ \nu_i^* \\ \epsilon_i^* \end{bmatrix}$$

where

$$\mathbf{B} = \begin{bmatrix} 1 & \frac{1}{T} & \cdots & \frac{1}{T} & \frac{1}{T} & \cdots & \frac{1}{T} \\ 0 & a_{01}^{\nu} - \mathbf{a}^{\mu} & \cdots & a_{0T}^{\nu} & a_{01}^{\epsilon} - \mathbf{a}^{\mu} & \cdots & a_{0T}^{\epsilon} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_{01}^{\nu} & \cdots & a_{0T}^{\nu} - \mathbf{a}^{\mu} & a_{01}^{\epsilon} & \cdots & a_{0T}^{\epsilon} - \mathbf{a}^{\mu} \\ 0 & a_{11}^{\nu} & \cdots & a_{1T}^{\nu} & \mathbf{a}^{\epsilon} & \cdots & a_{1T}^{\epsilon} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & a_{T1}^{\nu} & \cdots & a_{TT}^{\nu} & 0 & \cdots & \mathbf{a}^{\epsilon} \end{bmatrix}$$

Based on the assumption that the submatrix $\mathbf{B}_{[2:\text{end},2:\text{end}]}$ is invertible, we have

$$\frac{1}{T}\sum_t y_{it} = [1\ \frac{1}{T}\ \ldots\ \frac{1}{T}][\mu_i^*\ \nu_i^{*\prime}\ \epsilon_i^{*\prime}]' = [\frac{1}{T}\ \ldots\ \frac{1}{T}]\mathbf{B}_{[2:\text{end},2:\text{end}]}^{-1}[\delta_i^{\mu\prime}; \epsilon_i']' + \mu_i^*$$

Since $E([\nu_i^{*\prime}\ \epsilon_i^{*\prime}]'\mu_i^*) = \vec{0}$, and $[\delta_i^{\mu\prime}; \epsilon_i']'$ are linear combination of $[\nu_i^{*\prime}\ \epsilon_i^{*\prime}]'$, we have the moment conditions that

$$E([\delta_i^{\mu\prime}\ \epsilon_i']'\mu_i^*) = \vec{0} \tag{H1}$$

from which we could estimate $\mathbf{b}' \equiv [\frac{1}{T}\ \ldots\ \frac{1}{T}]\mathbf{B}_{[2:\text{end},2:\text{end}]}^{-1}$ and obtain

$$\hat{\mu}_i^* = \frac{1}{T}\sum_t y_{it} - \hat{\mathbf{b}}'[\delta_i^{\mu\prime}; \epsilon_i']'$$

Similarly, we obtain moments which allow us to estimate $\hat{\nu}_i^*$ and $\hat{\epsilon}_i^*$. Define $\delta_{it} \equiv \epsilon_{it} - \mathbf{a}^{\epsilon}(y_{it} - \mu_i^*)$, then we have

$$\begin{bmatrix} y_i - \mu_i^* \\ \delta_i \\ \epsilon_i \end{bmatrix} = \mathbf{C} \begin{bmatrix} \nu_i^* \\ \epsilon_i^* \end{bmatrix}$$

where

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 1 \\ a_{11}^{\nu} - \mathbf{a}^{\epsilon} & a_{12}^{\nu} & \cdots & a_{1T}^{\nu} & 0 & a_{12}^{\epsilon} & \cdots & a_{1T}^{\epsilon} \\ a_{21}^{\nu} & a_{22}^{\nu} - \mathbf{a}^{\epsilon} & \cdots & a_{2T}^{\nu} & 0 & 0 & \cdots & a_{2T}^{\epsilon} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{T1}^{\nu} & a_{T2}^{\nu} & \cdots & a_{TT}^{\nu} - \mathbf{a}^{\epsilon} & 0 & 0 & \cdots & 0 \\ a_{11}^{\nu} & a_{12}^{\nu} & \cdots & a_{1T}^{\nu} & \mathbf{a}^{\epsilon} & a_{12}^{\epsilon} & \cdots & a_{1T}^{\epsilon} \\ a_{21}^{\nu} & a_{22}^{\nu} & \cdots & a_{2T}^{\nu} & 0 & \mathbf{a}^{\epsilon} & \cdots & a_{2T}^{\epsilon} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{T1}^{\nu} & a_{T2}^{\nu} & \cdots & a_{TT}^{\nu} & 0 & 0 & \cdots & \mathbf{a}^{\epsilon} \end{bmatrix}$$

Starting from the first period, we have

$$y_{i1} - \mu_i^* = \nu_{i1}^* + \epsilon_{i1}^*$$

and it is easy to check that $[y_{i2} - \mu_i^* \ ... \ y_{iT} - \mu_i^* \ \sigma_{i1} \ \epsilon_{i2} \ ... \ \epsilon_{iT}]'$ is the result of a linear transformation of matrix $[\nu_{i1}^{*\prime} \ ... \ \nu_{iT}^{*\prime} \epsilon_{i2}^* \ ... \ \epsilon_{iT}^*]'$ with the coefficient $\mathbf{C_1}$. For simplicity of the notation, we define $M_{[-a,-b]}, a > 0, b > 0$ as the submatrix of M excluding row a and column b. Then matrix $\mathbf{C_1}$ can be expressed as $\mathbf{C_1} = [\mathbf{C}'_{[2:T,-(T+1)]} \ \mathbf{C}'_{[T+1,-(T+1)]} \ \mathbf{C}'_{[2T+2:\text{end},-(T+1)]}]'$.

With the assumption that $\mathbf{C_1}$ is invertible, we have the moment condition

$$E([y_{i2} - \mu_i^* \ ... \ y_{iT} - \mu_i^* \ \sigma_{i1} \ \epsilon_{i2} \ ... \ \epsilon_{iT}]'\epsilon_{i1}^*) = \vec{0} \tag{H2}$$

from which we could estimate $\mathbf{c_1}' \equiv [1 \ 0 \ ... \ 0]\mathbf{C_1}^{-1}$ and obtain

$$\hat{\epsilon}_{i1}^* = y_{i1} - \hat{\mu}_i^* - \hat{\mathbf{c}}_1'[y_{i2} - \hat{\mu}_i^* \ ... \ y_{iT} - \hat{\mu}_i^* \ \sigma_{i1} \ \epsilon_{i2} \ ... \ \epsilon_{iT}]'$$

Similarly, for any $t \in [1, ..., T]$, under the assumption that $\mathbf{C_t}$ is invertible, where

$$\mathbf{C_t} = [\mathbf{C}'_{[1:t-1,-(T+t)]} \ \mathbf{C}'_{[t+1:T,-(T+t)]} \ \mathbf{C}'_{[T+t,-(T+t)]} \ D'_{[1:t-1,-(T+t)]} \ \mathbf{C}'_{[2T+t+1:\text{end},-(T+t)]}]'$$

and $D \equiv [0_{T\times T} I_{T\times T}]$, we have moment condition

$$E([y_{i1} - \mu_i^* \ ... \ y_{i,t-1} - \mu_i^* \ y_{i,t+1} - \mu_i^* \ ... \ y_{iT} - \mu_i^* \ \sigma_{it} \ \epsilon_{i1}^* \ ...\epsilon_{i,t-1}^* \ \epsilon_{i,t+1} \ ... \ \epsilon_{iT}]'\epsilon_{it}^*) = \vec{0} \tag{H3}$$

Therefore, given any value of $\mathbf{a}^\mu$ and $\mathbf{a}^\epsilon$, we could estimate μ_i^* and ϵ_{it}^* through a sequential of linear regression as described later in this section.

Dynamics of ν_{it}^*. The model assumes that ν_{it}^* follows a first-order Markov process. This implies that the conditional distribution of ν_{it}^* given all the past information ν_i^{*t-1} only depends on the state of the previous period $\nu_{i,t-1}^*$, that is

$$f(\nu_{it}^*|\nu_i^{*t-1}) = f(\nu_{it}^*|\nu_{i,t-1}^*)$$

where $f(\cdot)$ is the density function. This further implies that the conditional mean of ν_{it}^* give all the past only depends on $\nu_{i,t-1}^*$:

$$E(\nu_{it}^*|\nu_i^{*t-1}) = E(\nu_{it}^*|\nu_{i,t-1}^*)$$

Therefore, we have

$$\nu_{it}^* = h(\nu_{i,t-1}^*) + e_{it}, E(e_{it}|\nu_i^{*t-1}) = 0$$

In fact we have infinite moment restrictions since $E(e_{it}r(v_i^{t-1})) = 0$ for any function $r(\cdot)$. Specifically, I use two sets of moments:

$$E(e_{it}h(v^*_{i,t-1})) = 0 \tag{H4}$$

$$E(e_{it}e_{ik}) = 0, \forall t \neq k \tag{H5}$$

From (*H4*) we obtain estimates of $\hat{h}(\cdot)$.

For simplicity, we do not solve the optimization jointly, instead, we solve the following constrainted optimization problem:

$$\min_{\mathbf{a}^\mu, \mathbf{a}^\epsilon} \sum_{t=1}^{T} \sum_{k=t+1}^{T} w_{tk} E(e_{it}e_{ik})^2$$

subject to conditions (H1), (H2), (H3), (H4). The optimization is managable since there are only two unknowns: $\mathbf{a}^\mu$ and $\mathbf{a}^\epsilon$.

In E-step, we obtain μ_i and v_{it} from posterior distribution $f_O(\mu_i, v_i | y_i; \hat{\Theta}^{(s)})$ and compute $\epsilon_{it} = y_{it} - \mu_i - v_{it}$. Now we explain in detail the M-step:

Step 1. Given each $\mathbf{a}^\mu$ and $\mathbf{a}^\epsilon$, obtain $\hat{\mathbf{A}}(\mathbf{a}^\mu, \mathbf{a}^\epsilon; \mu_i, v_i, \epsilon_i)$ and $\begin{bmatrix}\hat{\mu}_i^* & \hat{v}_i^{*\prime} & \hat{\epsilon}_i^{*\prime}\end{bmatrix}' = \hat{\mathbf{A}}^{-1}\begin{bmatrix}\mu_i & v_i' & \epsilon_i'\end{bmatrix}'$ using moments $\text{cov}(u_i^*, v_{it}^*) = 0, \text{cov}(u_i^*, \epsilon_{it}^*) = 0, \text{cov}(\epsilon_{ik}^*, \epsilon_{it}^*) = 0, \forall t, k \in [1, ..., T]$ and $t \neq k$

1. We first isolate $\hat{\mu}_i^*$ from y_{it}. Remember from matrix $\mathbf{A}$

   $$\mu_i = \mathbf{a}^\mu \mu_i^* + a_{01}^v v_{i1}^* + \cdots + a_{0T}^v v_{iT}^* + a_{01}^\epsilon \epsilon_{i1}^* + \cdots + a_{0T}^\epsilon \epsilon_{iT}^*$$

   and

   $$y_{it} = \mu_i^* + v_{it}^* + \epsilon_{it}^*$$

   For each $t \in [1, ..., T]$ we define

   $$\delta_{it}^\mu \equiv \mu_i - \mathbf{a}^\mu y_{it}$$

   Then δ_{it}^μ is a linear combination of $v_{i1}^*, ..., v_{iT}^*, \epsilon_{i1}^*, ..., \epsilon_{iT}^*$
   Finally, regress $\frac{1}{T}\sum_t y_{it}$ on $\delta_{i1}^\mu, ..., \delta_{iT}^\mu, \epsilon_{i1}, ..., \epsilon_{iT}$, and take the residual as estimates for μ_i^*, that is

   $$\hat{\mu}_i^* = \frac{1}{T}\sum_t y_{it} - X_i' \left(\sum_i X_i X_i'\right)^{-1} \left(\sum_i X_i (\frac{1}{T}\sum_t y_{it})\right)$$

   where $X_i \equiv [\delta_{i1}^\mu, ..., \delta_{iT}^\mu, \epsilon_{i1}, ..., \epsilon_{iT}]'$. By construction $\forall t \in [1, ..., T]$, δ_{it}^μ and ϵ_{it} are linear combinations of $v_{i1}^*, ..., v_{iT}^*, \epsilon_{i1}^*, ..., \epsilon_{iT}^*$. With assumption that μ_i^* is uncorrelated with v_i^* and ϵ_i^*, as well as extra assumption that $\delta_{i1}^\mu, ..., \delta_{iT}^\mu, \epsilon_{i1}, ..., \epsilon_{iT}$ are linearly independent, our estimator is able to isolate μ_i^* from y_{it}.

2. Next we isolate $\hat{v}^*_{it}$ from $y_{it} - \hat{\mu}^*_i$

   Complete the following loop:

   ***loop* for** $k = 1 : T$

   Define

   $$\delta_{ik} \equiv \epsilon_{ik} - \mathbf{a}^{\epsilon}(y_{ik} - \hat{\mu}^*_i)$$

   Then we can isolate $\hat{v}^*_{ik}$ and $\hat{\epsilon}^*_{ik}$ by

   $$\hat{\epsilon}^*_{ik} = y_{ik} - \hat{\mu}^*_i - S_i'\left(\sum_i S_i S_i'\right)^{-1}\left(\sum_i S_i\left(y_{ik} - \hat{\mu}^*_i\right)\right)$$

   and

   $$\hat{v}^*_{ik} = y_{ik} - \hat{\mu}^*_i - \hat{\epsilon}^*_{ik}$$

   where

   $$S_i = [y_{i1} - \hat{\mu}^*_i, ..., y_{i,k-1} - \hat{\mu}^*_i, y_{i,k+1} - \hat{\mu}^*_i, ..., y_{iT} - \hat{\mu}^*_i, \hat{\epsilon}^*_{i1}, ..., \hat{\epsilon}^*_{i,k-1}, \epsilon_{i,k+1}, ..., \epsilon_{i,T}, \delta_{ik}]'$$

   which is a linear transformation of $[v^*_{i1}, ..., v^*_{iT}, \epsilon^*_{i1}, ..., \epsilon^*_{i,k-1}, \epsilon^*_{i,k+1}, ..., \epsilon^*_{iT}]'$.

   Similar to previous step, $\forall k \in [1, ..., T]$, both elements of $[y_{i1} - \hat{\mu}^*_i, ..., y_{i,k-1} - \hat{\mu}^*_i, y_{i,k+1} - \hat{\mu}^*_i, ..., y_{iT} - \hat{\mu}^*_i, \hat{\epsilon}^*_{i1}, ..., \hat{\epsilon}^*_{i,k-1}, \epsilon_{i,k+1}, ..., \epsilon_{i,T}, \delta_{ik}]'$ and δ_{ik} are linear combinations of $v^*_{i1}, ..., v^*_{iT}, \epsilon^*_{i1}, ..., \epsilon^*_{i,k-1}, \epsilon^*_{i,k+1}, ..., \epsilon^*_{i,T}$. With assumption that ϵ^*_{ij} is uncorrelated with ϵ^*_{it} and v^*_i, for any $j \neq t$ and $j, t \in [1, ..., T]$, as well as extra assumption that $\hat{v}^*_{i1}, ..., \hat{v}^*_{i,k-1}, v_{i,k+1}, ..., v_{i,T}, \hat{\epsilon}^*_{i1}, ..., \hat{\epsilon}^*_{i,k-1}, \epsilon_{i,k+1}, ..., \epsilon_{i,T}$ and δ_{ik} are linearly independent, we could separate $\hat{v}^*_{it}$ and $\hat{\epsilon}^*_{it}$.

Step 2. Compute $\hat{\mathbf{a}}^{\mu}, \hat{\mathbf{a}}^{\epsilon} = \arg\min_{a^{\mu}, a^{\epsilon}} G\left(\hat{\mu}^*_i(a^{\mu}, a^{\epsilon}), \hat{v}^*_i(a^{\mu}, a^{\epsilon}), \hat{\epsilon}^*_i(a^{\mu}, a^{\epsilon})\right)$.

Function $G(\cdot)$ describes how well $\hat{\mu}^*_i(a^{\mu}, a^{\epsilon}), \hat{v}^*_i(a^{\mu}, a^{\epsilon}), \hat{\epsilon}^*_i(a^{\mu}, a^{\epsilon})$ satisfy other moment conditions. Specifically, we consider the following objects:

1. $v^*_{it}|v^*_{i,t-1}, v^*_{i,t-2}, ...$ follows first-order markov process $\Rightarrow$ regress v^*_{it} on $v^*_{i,t-2}, ..., v^*_{i,1}$ as well as polynomials of $v^*_{i,t-1}$, we should expect the coefficients on $v^*_{i,t-2}, ..., v^*_{i1}$ to be close to zero. Note here we only use the information on conditional mean, higher-order moments could also be exploited but should not at the expense of much longer computational time. For $t \in [3, ..., T]$,

   $$coef_t = \left(\sum_i res^r_i res^{r\prime}_i\right)^{-1}\left(\sum_i res^r_i res^v_{it}\right)$$

where

$$res_{it}^{v} = \hat{v}_{it}^{*} - P(\hat{v}_{i,t-1}^{*})'\left(\sum_i P(\hat{v}_{i,t-1}^{*})P(\hat{v}_{i,t-1}^{*})'\right)^{-1}\left(\sum_i P(\hat{v}_{i,t-1}^{*})\hat{v}_{it}^{*}\right)$$

$$res_{i}^{r} = \left([\hat{v}_{i,t-2}^{*},...,\hat{v}_{i1}^{*}] - P(\hat{v}_{i,t-1}^{*})'\left(\sum_i P(\hat{v}_{i,t-1}^{*})P(\hat{v}_{i,t-1}^{*})'\right)^{-1}\left(\sum_i P(\hat{v}_{i,t-1}^{*})[\hat{v}_{i,t-2}^{*},...,\hat{v}_{i1}^{*}]\right)\right)'$$

Function $P(x)$ is the low order Hermite polynomials of $g(x)$, which is a function of x[5]

Define

$$obj_1 = \left\|[coef_3', ..., coef_T']'\right\|$$

2. $v_{it}^{*}|v_{i,t-1}^{*}$ follows time-homogeneous markov process $\Rightarrow$ regress v_{it}^{*} on $P(v_{i,t-1}^{*})$ for any $t \in [2,...,T]$, we should expect the predicted conditional mean for different t given same v to be very close. For $t \in [2,...,T]$

$$\xi_t = \left(\sum_i P(\hat{v}_{i,t-1}^{*})P(\hat{v}_{i,t-1}^{*})'\right)^{-1}\left(\sum_i P(\hat{v}_{i,t-1}^{*})\hat{v}_{it}^{*}\right)$$

$$obj_2 = \left\|W^{\frac{1}{2}}\sum_j\left(P(\text{vec}(\hat{v}^{*}))\xi_j - \frac{1}{T-1}\sum_k P(\text{vec}(\hat{v}^{*}))\xi_k\right)^2\right\|$$

where W represents the histogram of $\hat{v}^{*}$.

3. ϵ_{it}^{*} i.i.d. over time $\Rightarrow \tau_{it}$ from uniform distribution which we know all the moments, where $\tau_{it} = \hat{Q}_{\epsilon}^{-1}(\epsilon_{it})$

$$obj_3 = \left\|\text{vec}\left(\frac{1}{N}[\vec{1}_{N\times 1} \;\; \tau \;\; \tau^2]' \times [\vec{1}_{N\times 1} \;\; \tau \;\; \tau^2]\right) - \text{vec}(M)\right\|$$

where M is the matrix of corresponding moments of joint uniform distribution.

Finally we solve the optimization

$$\hat{\mathbf{a}}^{\mu}, \hat{\mathbf{a}}^{\epsilon} = \arg\min_{a^{\mu}, a^{\epsilon}} k_1 obj_1 + k_2 obj_2 + k_3 obj_3$$

Comment: future experiment includes replacing current obj functions by a set of tests that take into account sampling errors

Step 3. Compute $\hat{\mu}_i^{*}(\hat{\mathbf{a}}^{\mu}, \hat{\mathbf{a}}^{\epsilon})$, $\hat{v}_i^{*}(\hat{\mathbf{a}}^{\mu}, \hat{\mathbf{a}}^{\epsilon})$, and $\hat{\epsilon}_i^{*}(\hat{\mathbf{a}}^{\mu}, \hat{\mathbf{a}}^{\epsilon})$, and update parameters by a series of quantile regressions

$$\hat{\gamma}_L^{Q}(\tau) = \arg\min_{\gamma_0^{Q}, \gamma_1^{Q}, ..., \gamma_K^{Q}} \sum_{i=1}^{N}\sum_{t=2}^{T}\rho_\tau\left(\hat{v}_{it}^{*} - \sum_{k=0}^{K}\gamma_k^{Q}\varphi_k(\hat{v}_{i,t-1}^{*})\right)$$

[5]In practice, we transform v_i using hyperbolic tangent function to reduce tails

$$\hat{\gamma}_L^{\epsilon}(\tau) = \arg\min_{\gamma^{\epsilon}} \sum_{i=1}^{N}\sum_{t=1}^{T} \rho_\tau\left(\hat{\epsilon}_{it}^* - \gamma^{\epsilon}\right)$$

$$\hat{\gamma}_L^{\nu_1}(\tau) = \arg\min_{\gamma^{\nu_1}} \sum_{i=1}^{N} \rho_\tau\left(\hat{\nu}_{i1}^* - \gamma^{\nu_1}\right)$$

$$\hat{\gamma}_L^{\mu}(\tau) = \arg\min_{\gamma^{\mu}} \sum_{i=1}^{N} \rho_\tau\left(\hat{\mu}_{i}^* - \gamma^{\mu}\right)$$

I Alternative PX-SEM for Quantile Models

We also consider a different way to implement PX-SEM for quantile models. The big structure is very similar to the previous method: try to recover $\hat{\mu}_i^*$, $\hat{\nu}_i^*$, and $\hat{\epsilon}_i^*$ from the E-step draws by imposing zero correlation, and then conduct quantile regressions based on $\hat{\mu}_i^*$, $\hat{\nu}_i^*$, and $\hat{\epsilon}_i^*$. The difference is that we now target matrix $\mathbf{A}^{-1}$ by writing down $\mathbf{A}^{-1}(\mathbf{a}^{\mu}, \mathbf{a}^{\epsilon})$, where $\mathbf{a}^{\mu}, \mathbf{a}^{\epsilon}$ are new auxiliary parameters. Here we only discuss how $\mathbf{A}^{-1}$ as well as $\hat{\mu}_i^*$, $\hat{\nu}_i^*$, and $\hat{\epsilon}_i^*$ are decided given $\mathbf{a}^{\mu}$ and $\mathbf{a}^{\epsilon}$, and the rest of the procedures of PX-SEM is the same as before.

Given any $\mathbf{a}^{\mu}$, $\mathbf{a}^{\epsilon}$, we get $[\hat{\mu}_i^* \;\; \hat{\nu}_i^* \;\; \hat{\epsilon}_i^*] = \mathbf{A}^{-1}[\mu_i \;\; \nu_i \;\; \epsilon_i]$ by imposing orthogonality

1. replace $\mu = \mu - X(pinv(x)\mu)$, $\nu_t = \nu_t + X(pinv(x)\mu)$ where $X = [1 \; \nu_1 \; ... \; \nu_T \; \epsilon_1 \; ... \; \epsilon_T]$

2. gen $\hat{\mu}^* \equiv \mathbf{a}^{\mu}\mu + x\frac{1}{T}\sum_t(y_t - \mu)$, and replace $\nu_t = y_t - \hat{\mu}^* - \epsilon_t$. Here given $\mathbf{a}^{\mu}$ and impose the orthogonality between $\hat{\mu}^*$ and $\frac{1}{T}\sum_t(y_t - \hat{\mu}_t^*)$, we can solve for $x = 0.5 \pm \sqrt{\mathbf{a}^{\mu}(1-\mathbf{a}^{\mu})\,\text{var}(\mu)/\text{var}(\frac{1}{T}\sum_t(y_t-\mu)) + 0.25}$ when $\mathbf{a}^{\mu} \neq 1, 0$. Choose the x that the correlation between $\hat{\mu}^*$ and $y - \hat{\mu}^*$ is closest to 0. When $\mathbf{a}^{\mu} = 1$ or 0 simply reg on other side

3. get $\hat{\epsilon}_t^*$ and $\hat{\nu}_t^*$. Iterate over t:

   ***loop* for $k = 1 : T$**

   (a) Reg ϵ_k on $[\hat{\nu}_1^* \; ... \; \hat{\nu}_{k-1}^* \; \hat{\nu}_{k+1}^* \; ... \; \hat{\nu}_T^* \; \hat{\epsilon}_1^* \; ... \; \hat{\epsilon}_{k-1}^* \; \epsilon_{k+1} \; ... \; \epsilon_T]$ and take residual res_k^{ϵ} ($\hat{\epsilon}_k$ which captures the correlation across periods will be added later to ν part)

   (b) Reg ν_k on $[\hat{\nu}_1^* \; ... \; \hat{\nu}_{k-1}^* \; \hat{\nu}_{k+1}^* \; ... \; \hat{\nu}_T^* \; \hat{\epsilon}_1^* \; ...\hat{\epsilon}_{k-1}^* \; res_k^{\epsilon} \; \epsilon_{k+1} \; ... \; \epsilon_T]$ and take residual res_k^{ν} ($\hat{\nu}_k$ which captures the correlation across periods will be

added later to v part)

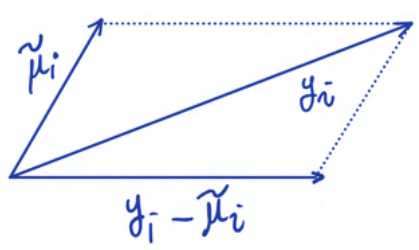
(c) Adjust the direction&length: gen $\hat{\epsilon}_k^* = \mathbf{a}^\epsilon res_k^\epsilon + x res_k^v$. Given $\mathbf{a}^\epsilon$ and orthogonality between res_k^ϵ and res_k^v, x can be solved (in step 2).

(d) Replace $\hat{v}_k^* = y - \hat{\mu}^* - \hat{\epsilon}_k^*$

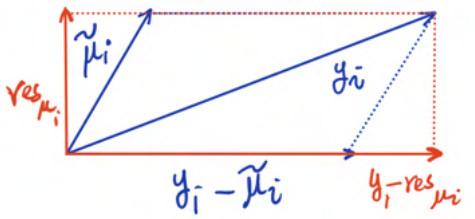
The procedure can be explained using the following figures. Assume there are only two variables. Figure I1a shows the draws from the E-step, where two variables are not orthogonal. Steps 3a and 3b aim to orthogonalize two variables and achieve the result as in Figure I1b. Then the auxiliary parameter is the length in current direction, and the finally vector is decided by imposing orthogonality as indicated in Figure I1c. One problem with this method is that we will have up to two possible vectors for a given value of the auxiliary parameter. We currently use one out of the two using the restrictions described above.

Figure I1: Procedures explained in figures

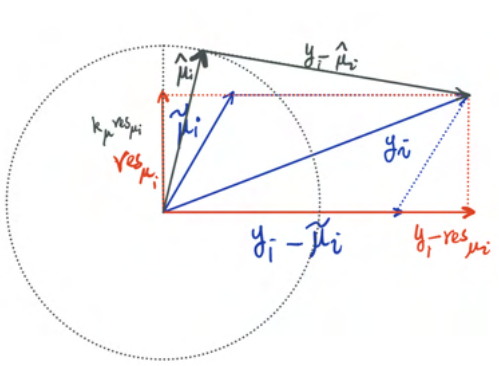
(a) original



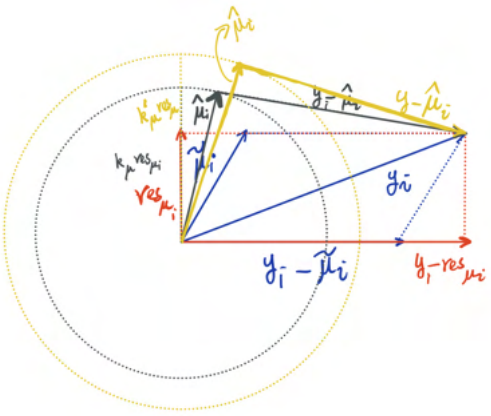
(b) orthogonalization



(c) scale by k and solve for A(k)



(d) scale by k' and solve for A(k')

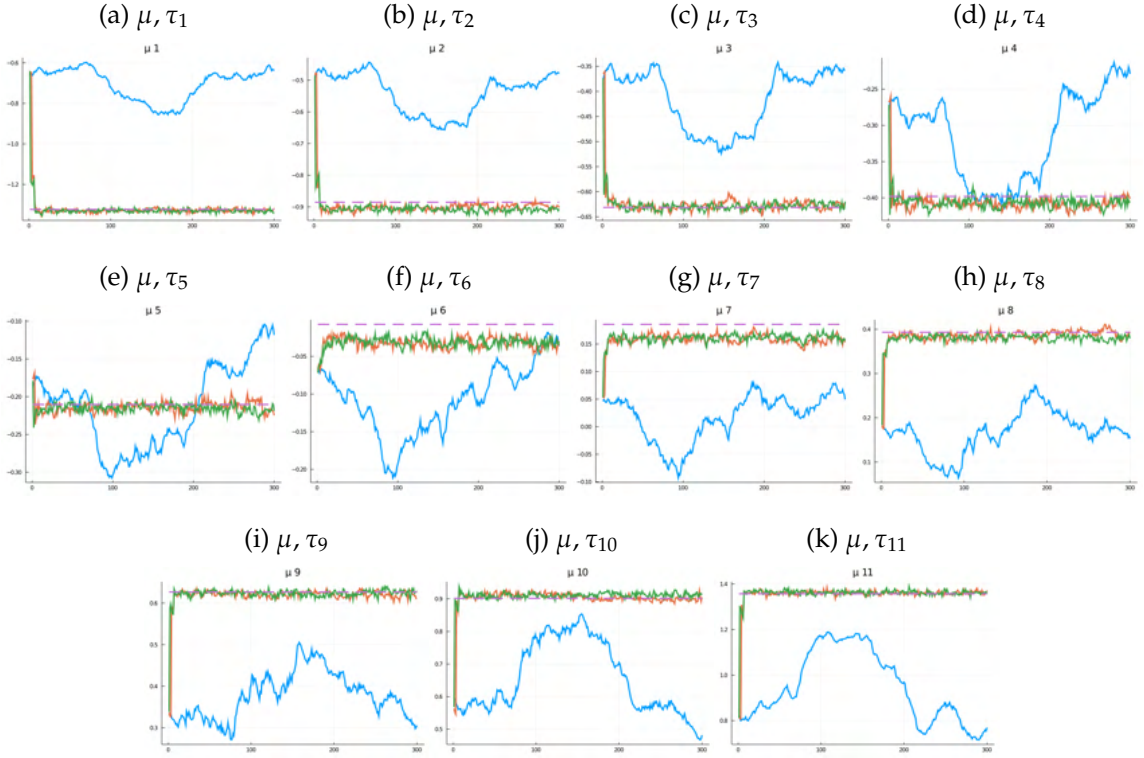


J Preliminary Results for Quantile Models

Here we present some simulation results of persistent-transitory quantile processes. Specifically, true μ_i and ϵ_i follow flexible non-Gaussian distribution, persistent compoenent $\nu_{it} = Q(\nu_{i,t-1}, \nu^2_{i,t-1}, u_{it})$ allows for nonlinear conditional mean and nonlinear persistence.

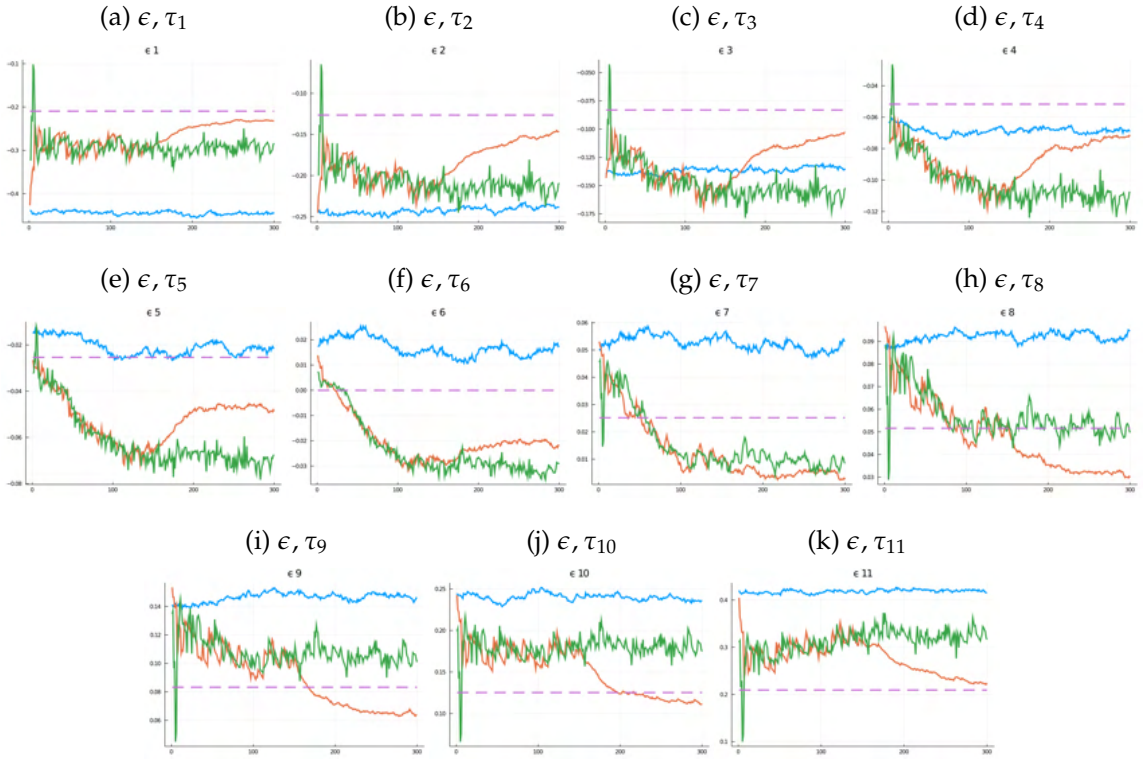
From the results shown below, we can see that SEM does not converge within 300 iterations, whereas both PX-SEM moves toward the region near the true value quickly. Based on the results of first 300 iterations, we plot the estimated distribution of each component. Not surprisingly, SEM does not capture the features of each component well. In contrast, both PX-SEM and PX-SEM+SEM, based on the first 300 iterations, could capture the nonlinear persistence associated with ν_{it}. Moreover, PX-SEM+SEM performs better in estimating the high kurtosis of ϵ_{it}, reflected in the iteration figures where the orange line moves closer to the true value at the second half.

Figure J1: SEM, PX-SEM, PX-SEM+SEM iterations of distribution parameters of μ_i



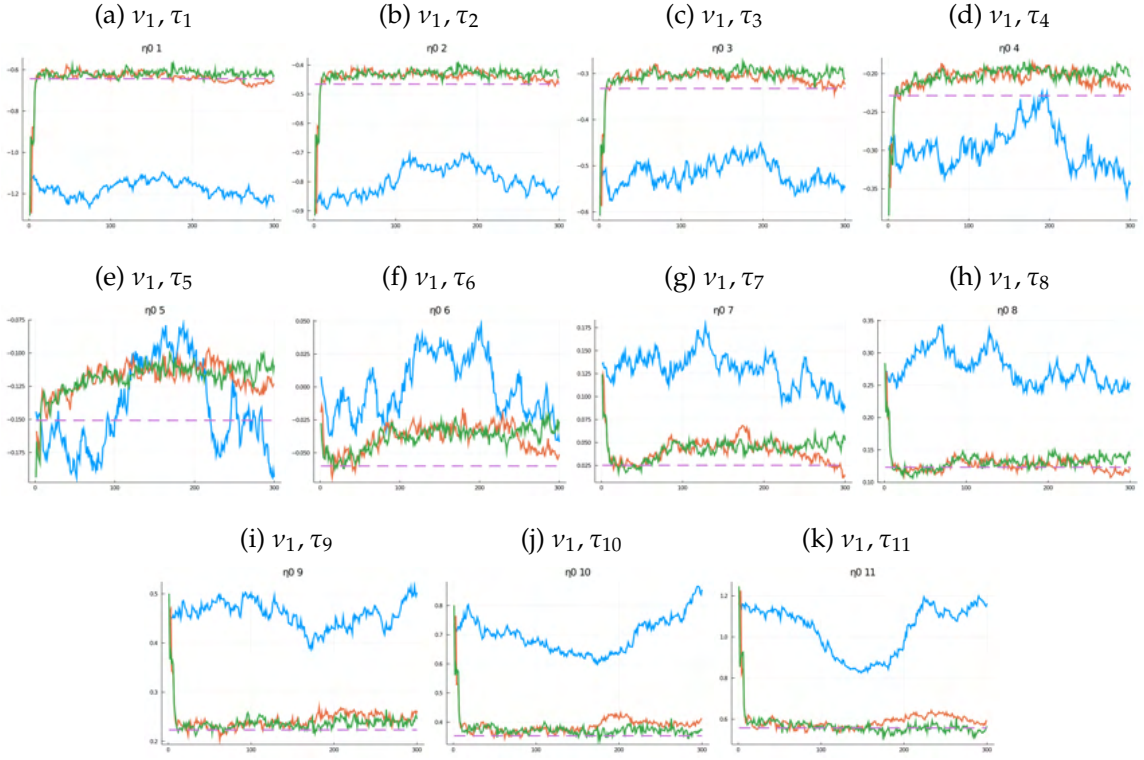
Notes: SEM—blue solid, PX-SEM — green solid, PX-SEM+SEM —orange solid, true values—pink dash.

Figure J2: SEM, PX-SEM, PX-SEM+SEM iterations of distribution parameters of ϵ_i



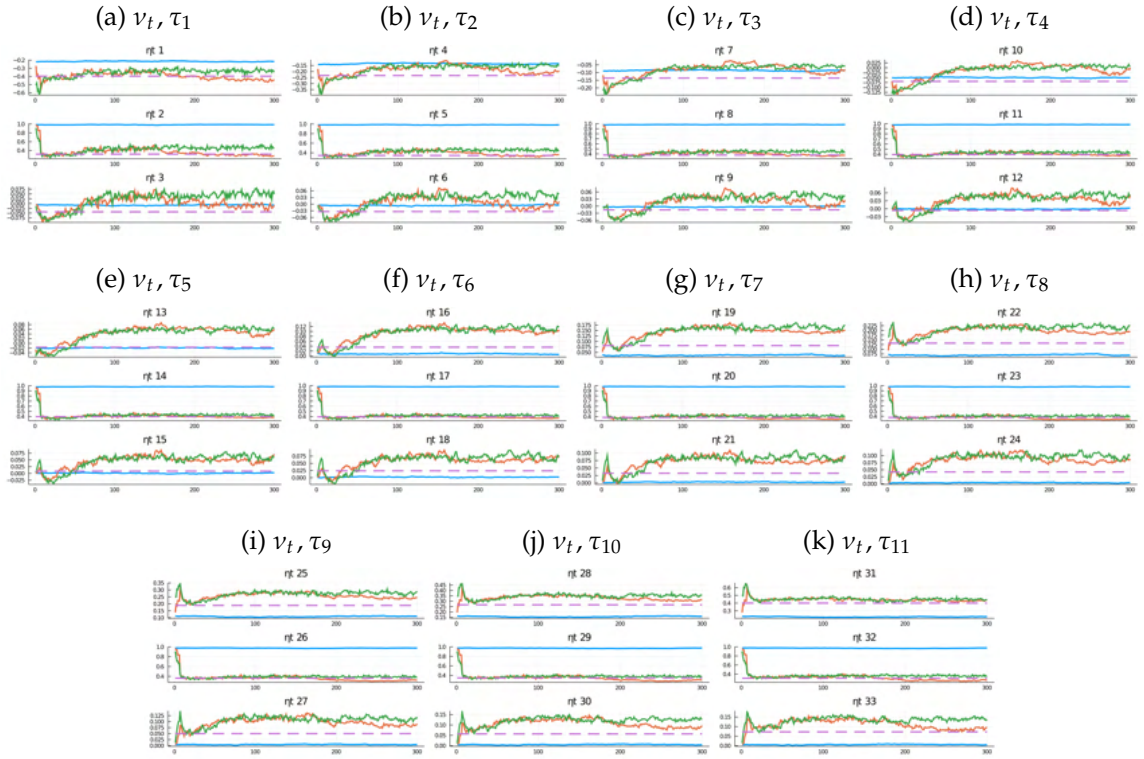
Notes: SEM—blue solid, PX-SEM — green solid, PX-SEM+SEM —orange solid, true values—pink dash.

Figure J3: SEM, PX-SEM, PX-SEM+SEM iterations of distribution parameters of ν_1



Notes: SEM—blue solid, PX-SEM — green solid, PX-SEM+SEM —orange solid, true values—pink dash.

Figure J4: SEM, PX-SEM, PX-SEM+SEM iterations of distribution parameters of $\nu_t|\nu_{t-1}$...



Notes: SEM—blue solid, PX-SEM — green solid, PX-SEM+SEM —orange solid, true values—pink dash.

Figure J5: SEM, PX-SEM, PX-SEM+SEM estimates of persistence of ν_t

(a) ν data (b) ν sem (c) ν pxsem (d) ν pxsem-sem

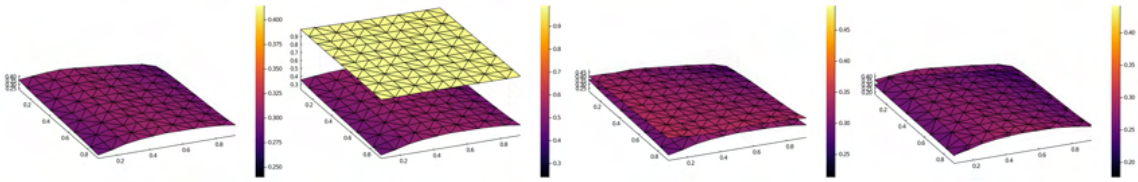
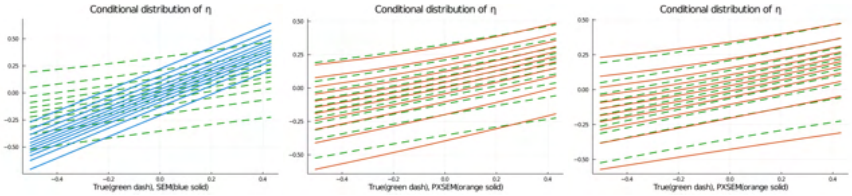


Figure J6: SEM, PX-SEM, PX-SEM+SEM estimates of conditional distribution of $\nu_t|\nu_{t-1}, \ldots$

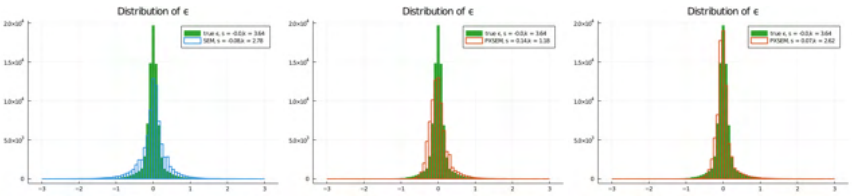
(a) ν sem (b) ν pxsem (c) ν pxsem-sem



Notes: SEM—blue solid, PX-SEM — orange solid, PX-SEM+SEM —orange solid, true values—green dash

Figure J7: SEM, PX-SEM, PX-SEM+SEM estimates of distribution of ϵ_i

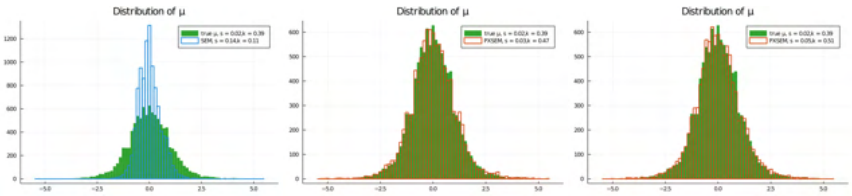
(a) ϵ sem (b) ϵ pxsem (c) ϵ pxsem-sem



Notes: SEM—blue histogram, PX-SEM — orange histogram, PX-SEM+SEM —orange histogram, true distribution—green histogram

Figure J8: SEM, PX-SEM, PX-SEM+SEM estimates of distribution of μ_i

(a) μ sem (b) μ pxsem (c) μ pxsem-sem



Notes: SEM—blue histogram, PX-SEM — orange histogram, PX-SEM+SEM —orange histogram, true distribution—green histogram