cemfi

working paper
2206

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

Siqi Wei

July 2022

CENTRO DE ESTUDIOS
MONETARIOS Y FINANCIEROS
Casado del Alisal 5, 28014 Madrid, Spain
www.cemfi.es

CEMFI Working Paper No. 2206
July 2022

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

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 high-dimensional 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 nonlinear panel models using the stochastic EM algorithm. We develop PX-SEM methods for two types of nonlinear panel data models: 1) binary choice models with individual effects and persistent shocks, and 2) persistent-transitory dynamic quantile processes. We find that PX-SEM can greatly speed up convergence especially when the initial guess is relatively far away from true values.

JEL Codes: C13, C33, C63.

Keywords: Stochastic EM, parameter-expansion, discrete choice model, dynamic quantile regression, latent variables.

Siqi Wei
CEMFI
siqi.wei@cemfi.edu.es

Acknowledgement

I am grateful to Manuel Arellano for his invaluable support and advice. I would also like to thank Martin Almuzara, Dmitry Arkhangelsky, Micole De Vera, Pedro Mira, Josep Pijoan-Mas, 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.

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 two types of nonlinear panel data models: 1) binary choice models with individual effects, persistent and transitory components and 2) persistent-transitory dynamic quantile processes.

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

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 two types of nonlinear panel data models, the binary choice models, and persistent-transitory dynamic quantile processes. The focus on nonlinear 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 both applications, we expand the O model linearly by allowing for a non-zero correlation between variables that are assumed to be non-correlated. Additionally, 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. In terms of estimation, we exploit moment-based estimators, which give us extra flexibility in relaxing 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 two types of nonlinear panel data models that are widely used in applied work: 1) discrete choice models and 2) 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 discusses general rules for conducting PX-SEM and explains the reason why the parameter expansion technique could speed up convergence. Next, I develop PX-SEM methods for two types of nonlinear latent-variable panel data models. In Section 4, I propose the PX-SEM algorithm for binary choice models, and in Section 5, I discuss the method for persistent-transitory dynamic quantile processes. Finally, Section 6 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:$\hat{\sigma} = \frac{1}{S^0} \sum_{S-S^0+1}^{S} \hat{\sigma}^{(s)}$.

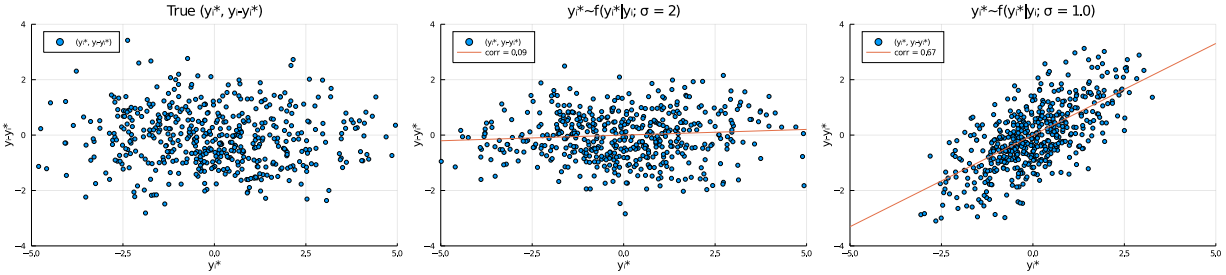
As described above, the two methods share the same E step, and the only difference is in the estimators of the M step: the PX-SEM estimator is adjusted by $\frac{1}{\hat{k}_L}$. Figure 1 explains what the PX-SEM and its auxiliary parameter k do. Specifically, Figure 1a is the scatter plot of simulations 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 $\sigma^{(s)} = \theta_{true}$, $(y_i^*, y_i - y_i^*)$ (c) Draws from E-step given $\sigma^{(s)} \neq \theta_{true}$, $(y_i^*, y_i - y_i^*)$



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 explain first the general steps for implementing the PX-SEM algorithm and then why it could improve the algorithmic efficiency.

Notation. Denote the model to be estimated, O Model, by $f_O(Y_i|X_i;\Theta) = \int_{Y_i^*} f_O(Y_i, Y_i^*|X_i;\Theta)dY_i^*$,

where Y_i and X_i are observable sets, Y_i^* is the latent-variable set, and Θ is the set of unknown parameters.

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 several restrictions: 1) L model nests O model: $\exists\ K_0$ such that $f_O(Y_i,Y_i^*|X_i;\Theta) = f_L(Y_i,Y_i^*|X_i;\Theta,K_0)$, and 2) There is a mapping from the L Model space to O Model space $\Theta = R(\Theta_L,K)$ such that the observed data is preserved $f_O(Y_i|X_i;R(\Theta_L,K)) = f_L(Y_i|X_i;\Theta_L,K)$

General steps. Under this setup, 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 $f_O(Y_i^*|Y_i,X_i;\hat{\Theta}^{(s)})$
2. PX-M step: Update parameters by
   (a) Estimate L model: $\hat{\Theta}_L, \hat{K} = \arg\min_{\Theta,K} \sum_i H(\Theta,K;Y_i,Y_i^*,X_i)$
   (b) Reduction: $\hat{\Theta}^{(s+1)} = R(\hat{\Theta}_L,\hat{K})$

where $H(\cdot)$ is a known function.[4]

Note that to implement PX-SEM, we do not need to put extra restrictions on the reduction function, which might give extra freedom in estimating the L model and is worthwhile as long as the reduction function can be obtained easily. However, in the following two sections where I develop PX-SEM algorithms for discrete choice models and quantile models, I *choose* the L model such that the reduction function satisfies $R(\Theta_L,K) = \Theta_L$: which means that the L model needs to satisfy $f_O(Y_i|X_i;\Theta_L) = f_L(Y_i|X_i;\Theta_L,K)$. Therefore, the E-step and PX-M steps become:

1. Stochastic E step: Draw Y_i^* from posterior distribution $f_O(Y_i^*|Y_i,X_i;\hat{\Theta}^{(s)})$
2. PX-M step: Update parameters by
   (a) Estimate L model: $\hat{\Theta}_L, \hat{K} = \arg\min_{\Theta,K} \sum_i H(\Theta,K;Y_i,Y_i^*,X_i)$
   (b) Reduction: $\hat{\Theta}^{(s+1)} = \hat{\Theta}_L$

Convergence. Liu et al., 1998 studies the convergence of $\hat{\Theta}^{(s)}$ based on a non-stochastic version, the PX-EM algorithm. Specifically, the paper proves that the likelihood of the observed-data model at each iteration increases, and discusses the conditions under

[4]For example, in case of MLE, $H(\Theta,K;Y_i,Y_i^*,X_i) = -\log f_L(Y_i,Y_i^*|X_i;\Theta,K)$.

which PX-EM dominates EM in the global rate of convergence. Here we briefly discuss it, which also helps understand the sources of potential gains in convergence.

It can be easily shown that:

$$\begin{aligned}
&\log f_O(Y_i|X_i;\hat{\Theta}^{(s+1)}) - \log f_O(Y_i|X_i;\hat{\Theta}^{(s)}) \\
&= \log f_L(Y_i|X_i;\hat{\Theta}_L,\hat{K}) - \log f_L(Y_i|X_i;\hat{\Theta}^{(s)},K_0) \\
&\geq Q(\hat{\Theta}_L,\hat{K}|\hat{\Theta}^{(s)},K_0) - Q(\hat{\Theta}^{(s)},K_0|\hat{\Theta}^{(s)},K_0) \geq 0
\end{aligned}$$

where $Q(\hat{\Theta}_L,\hat{K}|\hat{\Theta}^{(s)},K_0) = \int \log f_L(Y_i,Y_i^*|X_i;\hat{\Theta}_L,\hat{K}) f_L(Y_i^*|Y_i,X_i;\hat{\Theta}^{(s)},K_0) dY_i^*$. The first equation holds because of both assumption 1) L model nests O model at $K = K_0$, that is $f_O(Y_i|X_i;\hat{\Theta}^{(s)}) = f_O(Y_i|X_i;\hat{\Theta}^{(s)},K_0)$, and assumption 2) the reduction function exists and by construction $\hat{\Theta}^{(s+1)} = R(\hat{\Theta}_L,\hat{K})$, which implies $f_O(Y_i|X_i;\hat{\Theta}^{(s+1)}) = f_L(Y_i|X_i;\hat{\Theta}_L,\hat{K})$. The first inequality can be proved given Gibbs' inequality. And finally, the last inequality holds due to the definition of $\hat{\Theta}_L$: in case of the MLE, we have $\hat{\Theta}_L,\hat{K} = \arg\max \int \log f_L(Y_i,Y_i^*|X_i;\hat{\Theta}_L,\hat{K}) f_L(Y_i^*|Y_i,X_i;\hat{\Theta}^{(s)},K_0) dY_i^*$.[5]

From the equation above, we can easily see that, just like the EM algorithm, the PX-EM improves the observed-data likelihood in each iteration. On top of that, compared to the EM algorithm, the lower bound of the PX-EM increment is also larger because $Q(\hat{\Theta}_L,\hat{K}|\hat{\Theta}^{(s)},K_0) - Q(\hat{\Theta}^{(s)},K_0|\hat{\Theta}^{(s)},K_0) \geq Q(\hat{\Theta}_{SEM},K_o|\hat{\Theta}^{(s)},K_0) - Q(\hat{\Theta}^{(s)},K_0|\hat{\Theta}^{(s)},K_0)$ given that the L model is a more flexible one that nests the O model.

Statistical properties. Nielsen, 2000 studies the statistical properties of SEM algorithm based on the likelihood-based M-step. Specifically, the paper provides the conditions under which $\hat{\Theta}^{(s)}$ is ergodic, and has asymptotic normality

$$\sqrt{N}(\hat{\Theta}^{(s)} - \bar{\Theta}) \xrightarrow{d} N\left(0, I(\Theta_0)^{-1}(I - (I + F(\Theta_o))^{-1})\right)$$

given observables Y and X, and

$$\sqrt{N}(\hat{\Theta}^{(s)} - \Theta_0) \xrightarrow{d} N\left(0, I(\Theta_0)^{-1}(I - (2I + F(\Theta_o))^{-1})\right)$$

when $N \to \infty$.

Θ_0 is the true value, $\bar{\Theta}$ is the MLE $\frac{1}{N}\sum_{i=1}^N s_y(\bar{\Theta}) \equiv \frac{1}{N}\sum_{i=1}^N \frac{\partial}{\partial\Theta}\big|_{\bar{\Theta}} \log f_O(Y_i|X_i;\Theta) = 0$, $I(\Theta) = E(s_y(\Theta)s_y(\Theta)')$ is the observed data information, and $F(\theta) = I - I(\Theta)V(\Theta)^{-1}$ is the

[5] If in the PX-M step, we exploit other estimator rather than MLE, the last inequality may not hold. In fact, in Appendix B, we show an example when the GMM estimator exploited in M step combined with PX technique works worse than SEM with some specific initial guesses. However, when we use a larger L model, this is less of a concern.

expected fraction of missing information with $V(\Theta) = E(s_{y^*}(\Theta)s_{y^*}(\Theta)')$. $s_{y^*}(\Theta)$ is defined as $\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\Theta}\big|_{\bar{\Theta}} \log f_O(Y_i, Y|X_i;\Theta)$.

Additionally, as proved in Liu et al., 1998, $V_L(\Theta)^{-1} \geq V_O(\Theta)^{-1}$ in semipositive definite order. Then it is easy to show that $I(\Theta_0)^{-1}(I-(I+F_L(\Theta_o))^{-1}) \leq I(\Theta_0)^{-1}(I-(I+F_O(\Theta_o))^{-1})$ in semipositive definite order, that is saying that not only converging faster, likelihood-based PX-SEM estimator is also more efficient than standard SEM estimator.[6]

A formal discussion of the large sample properties of moments-based PX-SEM is beyond the scope of this chapter. In practice, we usually find that it is less efficient than estimators based on a standard SEM algorithm. However, we might still want to use the moments-based PX-SEM estimator for at least two reasons. First, the efficiency is not a big concern if we 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. Secondly, there might be a tradeoff between the flexibility of the L model and PX-M step estimators. Specifically, allowing for a moments-based PX-M step estimator might give us extra flexibility in relaxing the L model, which could deliver much faster convergence. For example, in Appendix B, we show an example when the moments-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 two sections, we will discuss two examples: 1) discrete choice models and 2) quantile models.

[6] To prove $I(\Theta_0)^{-1}(I-(I+F_L(\Theta_o))^{-1}) - I(\Theta_0)^{-1}(I-(I+F_O(\Theta_o))^{-1}) = I(\Theta_0)^{-1}(I+F_O(\Theta_0))^{-1} - I(\Theta_0)^{-1}(I+F_L(\Theta_0))^{-1} \leq 0$, equivalent to proving $(I+F_O(\Theta_0))I(\Theta_0) - (I+F_L(\Theta_0))I(\Theta_0) = I(\Theta_0)(V_L(\Theta_0)^{-1} - V_O(\Theta_0)^{-1})I(\Theta_0) \geq 0$

4 Discrete Choice Models

The first 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:

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

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

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

where $\mu_i | x \sim N(0, \sigma_\mu^2)$, $u_{it} \sim N(0, \sigma_u^2)$, $\epsilon_{it} \sim N(0,1)$, $\nu_{i1} \sim N(0,1)$.[7]

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 $\Theta = (\beta, \sigma_\mu, \rho, \sigma_u)$.

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 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. M step:
   - $\hat{\beta}^{(s+1)} = (\sum x_{it} x_{it}')^{-1} (\sum x_{it}(y_{it}^* - \mu_i - \nu_{it}))$
   - $\hat{\sigma}_\mu^{(s+1)} = \widehat{\text{std}}(\mu_i^*)$
   - $\hat{\rho}^{(s+1)} = (\sum \nu_{i,t-1} \nu_{i,t-1}')^{-1} (\sum \nu_{i,t-1} \nu_{it})$
   - $\hat{\sigma}_u^{(s+1)} = \widehat{\text{std}}(\nu_{it} - \hat{\rho} \nu_{i,t-1})$

[7] In more flexible model, estimate var(ν_{i1}) as well.

PX-SEM algorithm. To implement PX-SEM algorithm, we propose the following L model:

L Model

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

$$y^*_{it} = \gamma_t' x_i + \mu_i + \nu_{it} + \epsilon_{it},$$

$$\begin{bmatrix} \mu_i \\ \nu_i \\ \epsilon_i \end{bmatrix} \Bigg| x_i \sim N\Bigg(\mathbf{B}x_i,\ \mathbf{p}^2\mathbf{A}\underbrace{\begin{bmatrix} \sigma^2_\mu & 0 & 0 \\ 0 & \Sigma_\nu(\rho,\sigma_u) & 0 \\ 0 & 0 & I \end{bmatrix}}_{\Sigma}\mathbf{A}'\Bigg), \mathbf{p} > 0$$

where $x_i = [x_{i1}; ...; x_{iT}]$, $\mathbf{B} = \begin{bmatrix} b'_\mu \\ b'_\nu \\ b'_\epsilon \end{bmatrix}$, subject to $\beta = \frac{1}{\mathbf{p}}(\gamma_t + b_\mu + b_{\nu,t} + b_{\epsilon,t})$, $C\mathbf{A}\Sigma\mathbf{A}'C' = C\Sigma C'$, $C = [\overrightarrow{1}_{T\times 1}\ I_{T\times T}\ I_{T\times T}]$, and $\mathbf{A}$ is a lower triangular matrix with postive diagonal entries. Matrix $\Sigma_\nu(\rho,\sigma_u)$ is the variance-covariance matrix of vector ν_i which follows the AR(1) structure.

This L model contains three types of auxiliary parameters to deal with three different violations of model assumptions that might happen to the draws under some parameter guesses. Matrix $\mathbf{B}$ takes care of the linear correlation between observables x_i and latent draws (e.g., the potential positive correlation when the guess $\hat{\beta}$ in E-step is much smaller than the true value in absolute level, to account for the correlation between x_i and y_i); Matrix $\mathbf{A}$ allows for linear correlation among μ_i, ν_1, u_{it} and ϵ_i (e.g., the potential positive correlation among u_i across periods when the guess σ_μ is too small, to account for the correlation between $y_{i,t}$ and $y_{i,k}$); Scalar $\mathbf{p}$ scales up and down y^*_{it} and allows the var(ϵ_{it}) to be different from 1 (e.g., when the guess of other parameters too small). It is easy to verify that L model nests the O model: when $\mathbf{B} = \overrightarrow{0}_{(2T+1)\times K}$, $\mathbf{A} = I_{(2T+1)\times(2T+1)}$, $\mathbf{p} = 1$, the two models coincide $f_O(y_i, y^*_i, \mu_i, \nu_i|x_i;\Theta) = f_L(y_i, y^*_i, \mu_i, \nu_i|x_i;\Theta, \mathbf{A} = I, \mathbf{B} = \overrightarrow{0}, \mathbf{p} = 1)$.

Additionally, constraints $\beta = \frac{1}{\mathbf{p}}(\gamma_t + b_\mu + b_{\nu,t} + b_{\epsilon,t})$ and $C\mathbf{A}\Sigma\mathbf{A}'C' = C\Sigma C'$ guarantee that $f_O(y_i|x_i;\Theta) = f_L(y_i|x_i;\Theta, \mathbf{A}, \mathbf{B}, \mathbf{p})$ for any $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{p}$. These are not necessary conditions to implement PX-SEM algorithm. However, it simplifies the mapping from the L model space to O model space, as it becomes identity mapping $\hat{\Theta}^{(s+1)} = R(\hat{\Theta}^{(s+1)}_L, \mathbf{A}^{(s+1)}, \mathbf{B}^{(s+1)}, \mathbf{p}^{(s+1)}) = \hat{\Theta}^{(s+1)}_L$.

Finally, we discuss the procedures to implement PX-SEM algorithm. Similarly, 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. $\hat{\sigma}_\mu, \hat{\mathbf{p}} = \arg\min_{\sigma_\mu, p} \left\| W^{\frac{1}{2}} \left(\text{vec}(\hat{\Sigma}_\nu) - \text{vec}(\Sigma_\nu(\rho(\sigma_\mu, p), \sigma_u(\sigma_\mu, p))) \right) \right\|$, where $\hat{\Sigma}_\nu = \widehat{\text{cov}}(\tilde{y}_i) - \sigma_\mu^2 p^2 \vec{1}_{T\times 1} \vec{1}_{1\times T} - p^2 I_{T\times T}$, $\rho(\sigma_\mu, p) = (\sum \tilde{\tilde{\nu}}_{i,t-1} \tilde{\tilde{\nu}}_{i,t-1}')^{-1} (\sum \tilde{\tilde{\nu}}_{i,t-1} \tilde{\tilde{\nu}}_{i,t})$, $\sigma_u(\sigma_\mu, p) = \widehat{\text{std}}(\tilde{\tilde{\nu}}_{it} - \rho \tilde{\tilde{\nu}}_{i,t-1})$, and $\tilde{\tilde{\nu}}_i = \text{chol}(\hat{\Sigma}_\nu) \, \text{chol}(\widehat{\text{cov}}([\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i]))^{-1} [\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i]$
      iv. $\hat{\rho} = (\sum \tilde{\tilde{\nu}}_{i,t-1} \tilde{\tilde{\nu}}_{i,t-1}')^{-1} (\sum \tilde{\tilde{\nu}}_{i,t-1} \tilde{\tilde{\nu}}_{i,t})$, $\hat{\sigma}_u = \widehat{\text{std}}(\tilde{\tilde{\nu}}_{it} - \hat{\rho} \tilde{\tilde{\nu}}_{i,t-1})$, $\hat{\beta} = \frac{\widehat{\mathbf{p}\beta}}{\hat{\mathbf{p}}}$, where $\tilde{\tilde{\nu}}_i = \text{chol}(\hat{\Sigma}_\nu(\hat{\sigma}_\mu, \hat{\mathbf{p}})) \, \text{chol}(\widehat{\text{cov}}([\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i]))^{-1} [\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i]$
      v. $\hat{\Theta}_L = (\hat{\beta}, \hat{\sigma}_\mu, \hat{\rho}, \hat{\sigma}_u)$
   (b) Reduction: identity mapping due to L model constraints $\hat{\Theta}^{(s+1)} = \hat{\Theta}_L$

Intuitively speaking, this L model assumes that the latent variables are possibly some affine transformation of "original" latent variables in the O model. Therefore, the objective of the M-step is to figure out the affine mapping, which consists of auxiliary parameters, of latent draws in E-step whose outcome satisfy constraints of the O model.

Specifically, the estimator $\hat{\sigma}_\mu$ and $\hat{\mathbf{p}}$ in step 2(a)iii are obtained through the following two constraints:

- $p^2 C \widehat{\mathbf{A}\Sigma\mathbf{A}'} C' = C \widehat{\text{cov}}([\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i]) C'$, $C\mathbf{A}\Sigma\mathbf{A}'C' = C\Sigma C' \Rightarrow C \widehat{\text{cov}}([\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i]) C' = p^2 C \Sigma C' \Rightarrow$ write down Σ as a function of p and σ_μ and specifically, $\hat{\Sigma}_\nu = \widehat{\text{cov}}(\tilde{y}_i) - \sigma_\mu^2 p^2 \vec{1}_{T\times 1} \vec{1}_{1\times T} - p^2 I_{T\times T}$
- $\text{cov}([\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i]) = p^2 A \Sigma A' \Rightarrow \widehat{pA} = \text{chol}(\widehat{\text{cov}}([\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i])) \, \text{chol}(\Sigma)^{-1}$, and $\widehat{pA}^{-1} [\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i] \sim N(0, \Sigma)$, specifically $\widehat{pA}^{-1}_{(2:T+1,:)} [\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i]$ part follows AR(1) $\Rightarrow$ estimate of ρ and σ_u using $\widehat{pA}^{-1}_{(2:T+1,:)} [\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i]$ and construct $\Sigma_\nu(\rho, \sigma_u)$ using AR(1) structure

Estimators $\hat{\sigma}_\mu$ and $\hat{\mathbf{p}}$ are chosen to minimize the distance between the two matrix.[8,9]

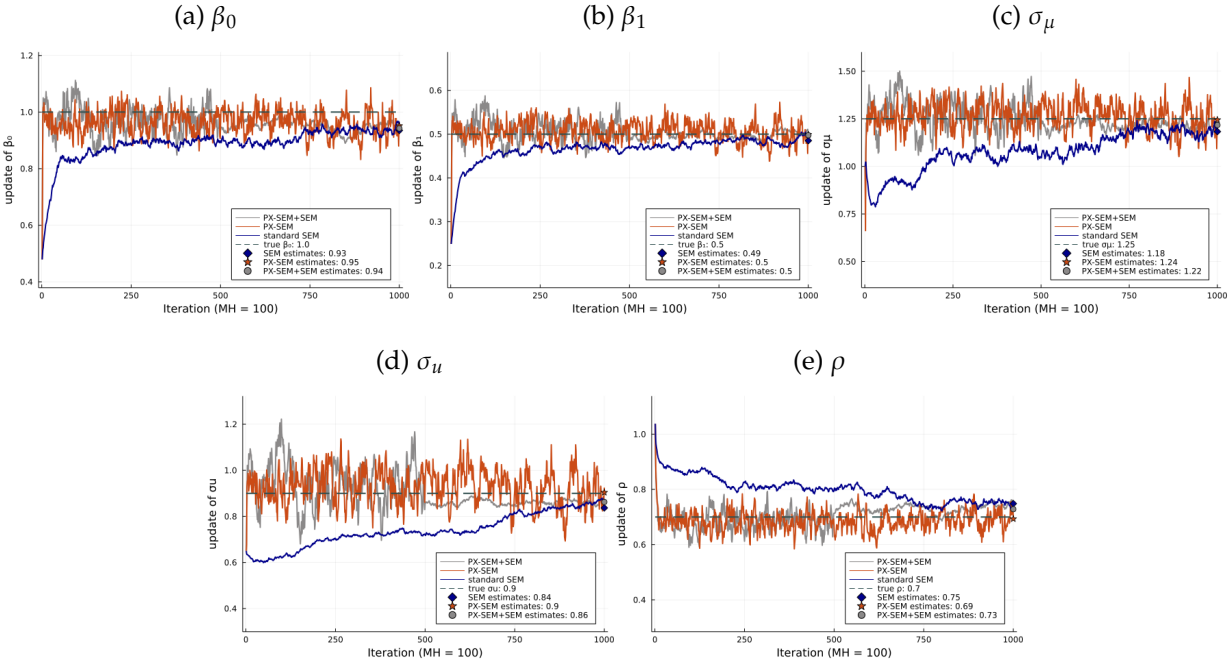
[8] Different estimators of L model can be exploited as well. For example joint optimization: $\hat{\sigma}_\mu, \hat{\mathbf{p}}, \hat{\rho}, \hat{\sigma}_u = \arg\min_{\sigma_\mu, p, \rho, \sigma_u} \left\| W^{\frac{1}{2}} \left(\text{vec}(C \widehat{\text{cov}}([\tilde{\mu}_i; \tilde{\nu}_i; \tilde{\epsilon}_i])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

[9] 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.

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 2: 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.

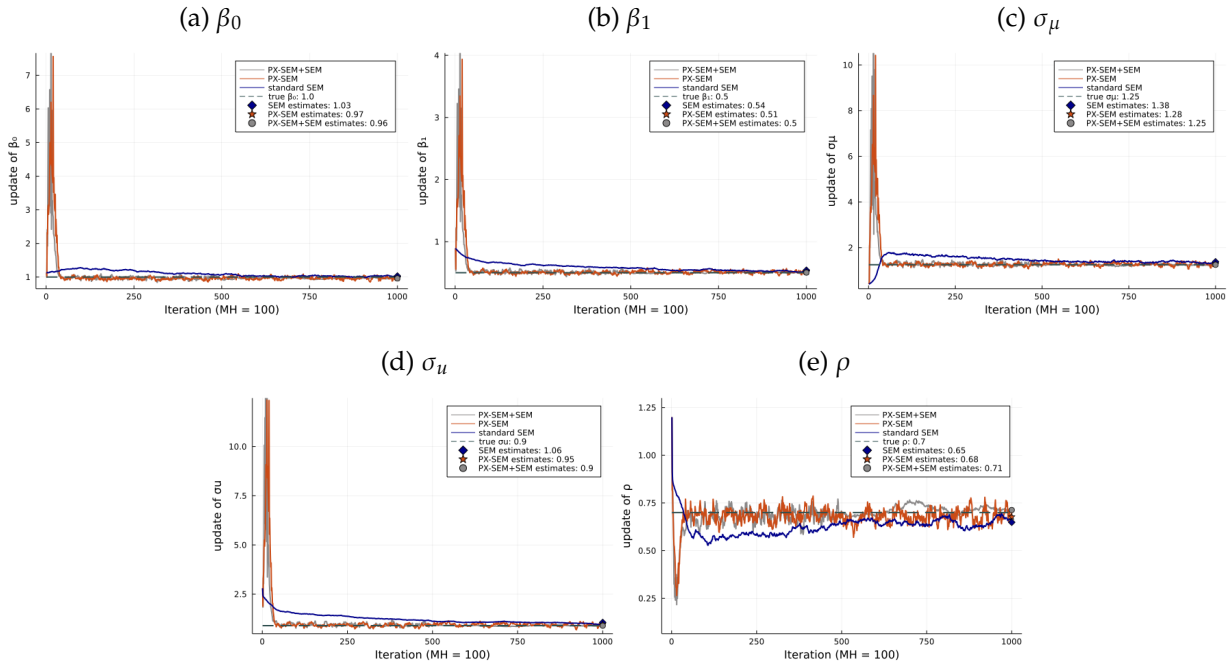
Figure 2 presents the estimation results of one simulation with $N = 5000$ and $T = 8$. Specifically, we plot the M-step updates $\hat{\Theta}^{(s)}$ for 1000 iterations ($S = 1000$). The blue line shows each update of the SEM algorithm, whereas the orange one represents the PX-SEM updates. We also combine 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

[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 $\hat{\rho}^{(0)} = \frac{1}{N(T-1)} \sum \frac{\widehat{res}_{i,t-1}\widehat{res}_{it}}{\widehat{res}_{i,t-1}^2}$, and $\hat{\sigma}_u^{(0)} = \widehat{\text{std}}(\widehat{res}_{it} - \hat{\rho}^{(0)}\widehat{res}_{i,t-1})$

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.

Figure 3: SEM and PX-SEM iterations of $\Theta^{(s)}$ from random initial guesses



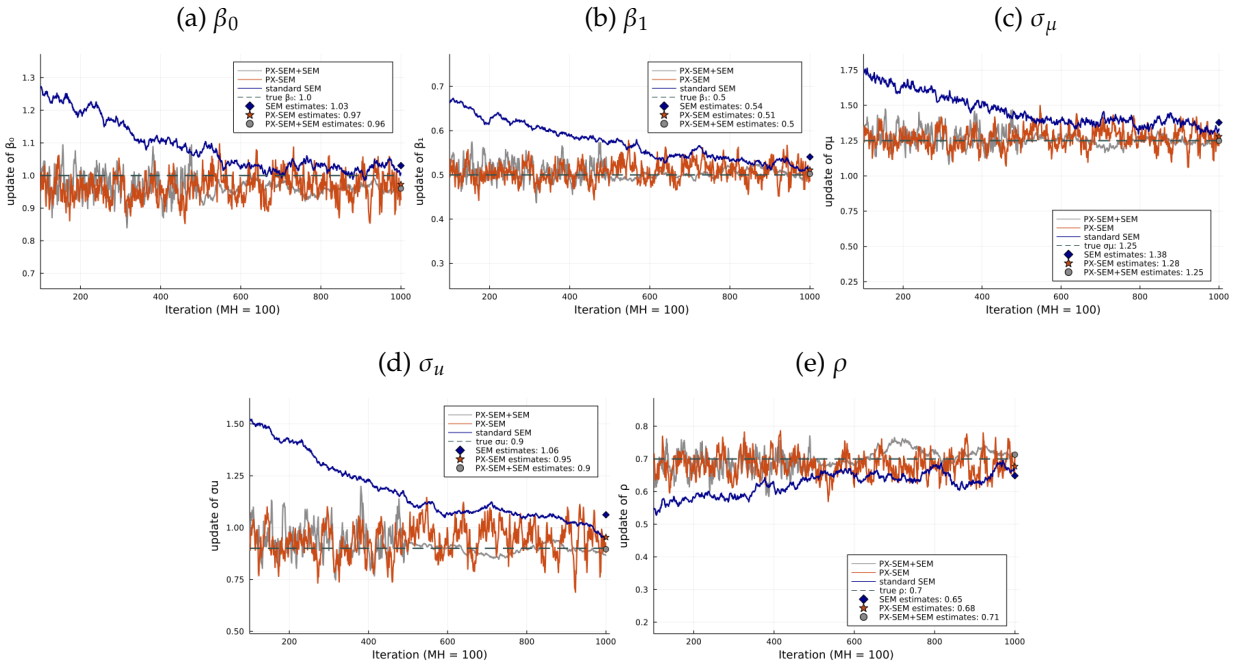
Notes: Complete iterations of SEM (blue solid line), PX-SEM (orange solid line), and PX-SEM + SEM(grey solid line, 500 iterations each) based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond), PX-SEM estimates (orange star), and PX-SEM+SEM (grey circle) are all based on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

In Figure 3, 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,

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

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.

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



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 3.

To have a better view of the details, we plot the last 900 iterations in Figure 4. 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 D.

5 Quantile Models

The second 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.

O Model:

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

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

where ϵ_{it} has zero mean, i.i.d. over time, and independent of ν_i and μ_i.

Unlike the canonical permanent-transitory process, this model allows for a much more flexible conditional distribution of persistent component ν_{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:

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

$$Q_\epsilon(\tau) = \gamma^\epsilon(\tau), \quad Q_{\nu_1}(\tau) = \gamma^{\nu_1}(\tau), \quad Q_\mu(\tau) = \gamma^\mu(\tau)$$

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 $\gamma_k^Q(\tau)$, $\gamma^\epsilon(\tau)$, $\gamma^{\nu_1}(\tau)$, 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:

1. Stochastic E step: Draw μ_i and ν_i from posterior distribution $f_O(\mu_i, \nu_i|y_i; \hat{\Theta}^{(s)})$

[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 ν_{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.

2. M step: Update parameters by computing a series of quantile regressions:

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

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

$$\hat{\gamma}^{v_1}(\tau) = \arg \min_{\gamma^{v_1}} \sum_{i=1}^{N} \rho_\tau \left(v_{i1} - \gamma^{v_1} \right)$$

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

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

L Model:

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

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

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

where ϵ_{it}^* has zero mean, i.i.d. over time, and independent of v_i^* and μ_i^*. Auxiliary matrix $\mathbf{A}$ helps to expand linearly the O model. Additionally, to simplify the "reduction" process, we restrict $f_O(y_i; \Theta) = f_L(y_i; \Theta, \mathbf{A})$ by imposing $C\mathbf{A} = C$, $C = [\overrightarrow{1}_{T\times 1} \; I_{T\times T} \; I_{T\times T}]$. Compared to the discrete choice model case, this is more restrictive as it put restrictions on each observation $y_{it} = \mu_i^* + v_{it}^* + \epsilon_{it}^* = \mu_i + v_{it} + \epsilon_{it}$ instead of the joint distribution of y_i.[14] Finally, it is to show that when $\mathbf{A} = I_{T\times T}$, L model coincides with O model.

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 v_i from posterior distribution $f_O(\mu_i, v_i | y_i; \hat{\Theta}^{(s)})$
2. PX-M step:
   (a) Estimate L model: $\hat{\Theta}_L$, $\hat{\mathbf{A}}$
   (b) Reduction: $\hat{\Theta}^{(s+1)} = \hat{\Theta}_L$ (due to constraints $C\mathbf{A} = C$)

[14] In the discrete choice model case, because of the assumption of normality, we only need to target the first two moments to guarantee $f_O(y_i; \Theta) = f_L(y_i; \Theta, \mathbf{A})$.

The E-step is the same as the SEM algorithm, which is drawing μ_i and ν_{it} from the posterior distribution $f_O(\mu_i, \nu_i | y_i; \hat{\Theta}^{(s)})$. However, the difficulty lies in the estimation of the L model. With the complication of matrix $\mathbf{A}$, we can no longer simply conduct a series of quantile regressions directly using E-step draws as in the SEM. Considering that the number of unknown parameters is not small, we will estimate parameters sequentially. First, we obtain estimates $\hat{\mathbf{A}}$, 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 moment conditions, including zero correlations among μ_i^*, ν_i^*, and ϵ_i^*, and ν_{it}^* following a time-homogeneous first-order Markov process. Secondly, we estimate the rest of the parameters including $\hat{\gamma}$ by a series of quantile regressions of $\hat{\mu}_i^*$, $\hat{\nu}_i^*$, and $\hat{\epsilon}_i^*$.

Now we discuss in detail one potential way of implementing PX-SEM in this example. We specify matrix $\mathbf{A}$ as follows:

$$\mathbf{A} = \left[\begin{array}{cccccccc}
\mathbf{a}^\mu & a_{01}^\nu & \cdots & a_{0T}^\nu & a_{01}^\epsilon & \cdots & a_{0T}^\epsilon \\
\hline
1-a^\mu & 1-a_{o1}^\nu-a_{11}^\nu & \cdots & -a_{0T}^\nu-a_{1T}^\nu & 1-a_{o1}^\epsilon-\mathbf{a}^\epsilon & \cdots & -a_{0T}^\epsilon-a_{1T}^\epsilon \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
1-a^\mu & -a_{01}^\nu-a_{T1}^\nu & \cdots & 1-a_{0T}^\nu-a_{TT}^\nu & 0 & \cdots & 1-a_{oT}^\epsilon-\mathbf{a}^\epsilon \\
\hline
0 & a_{11}^\nu & \cdots & a_{1T}^\nu & \mathbf{a}^\epsilon & \cdots & a_{1T}^\epsilon \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & -a_{T1}^\nu & \cdots & a_{TT}^\nu & 0 & \cdots & \mathbf{a}^\epsilon
\end{array}\right] \begin{array}{l} \} \mathbf{A1}_{1\times(2T+1)} \\ \\ \} \mathbf{A2}_{T\times(2T+1)} \\ \\ \} \mathbf{A3}_{T\times(2T+1)} \end{array}$$

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

Now we explain how to estimate L model:

1. 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

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

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

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

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

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

Some preliminary simulation results are shown in Appendix G. 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 which include: 1) numerical stability associated with nonlinear optimization, 2) optimization does not take into account sampling errors which might be an issue with small sample size, and 3) the choice of weighting matrix.

6 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 two types of models: 1) discrete choice models with individual effects, persistent and transitory components, and 2) 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.

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.

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.

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:

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

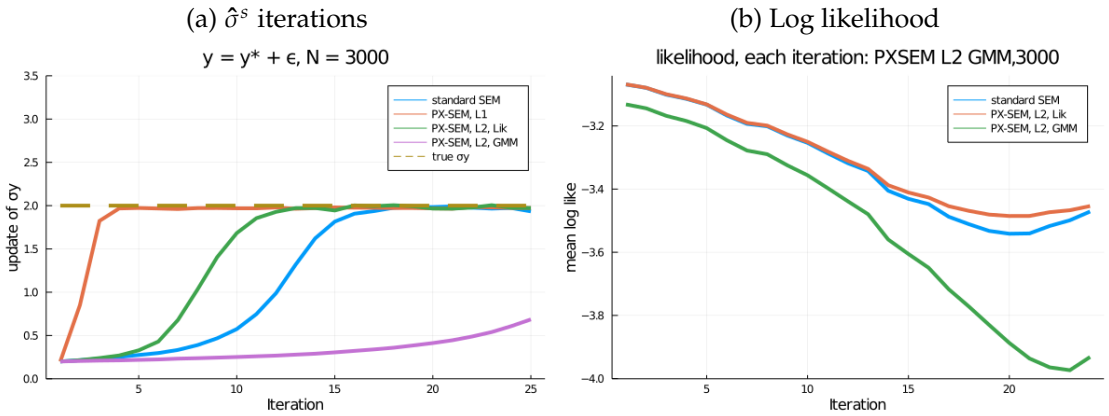
where $[y_i^* \ \ \epsilon_i]' \sim N(0, \begin{pmatrix} 1 & 1-k \\ 0 & k \end{pmatrix}\begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1-k \\ 0 & k \end{pmatrix}') \Leftrightarrow \text{N}(0, \begin{pmatrix} \sigma^2 + (1-k)^2 & k(1-k) \\ k(1-k) & k^2 \end{pmatrix})$

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



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* $s = 0, ..., S$

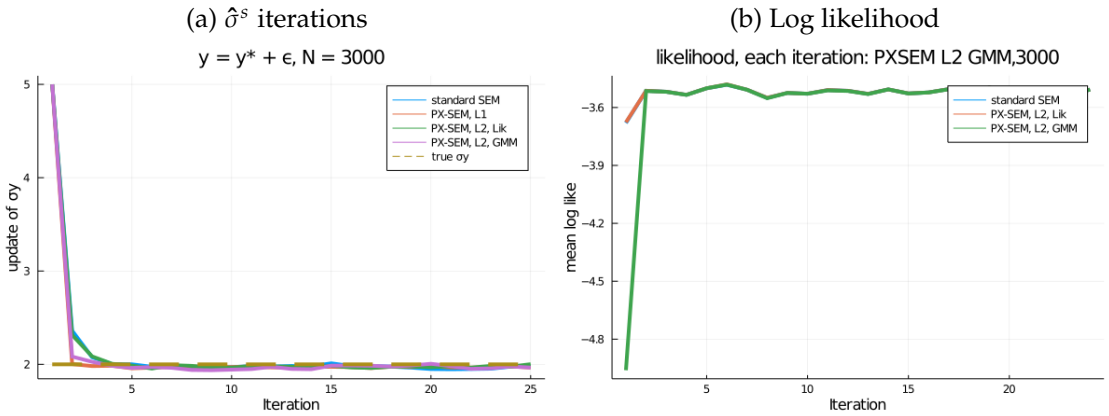
Then we compare SEM (blue), PX-SEM-L2-1 (purple), and PX-SEM-L2-2 (green). As we can see in the figure, PX-SEM-L2-1, with GMM based PX-M-step, performs even worse than SEM. This is not surprising though. It is easy to see why it is happening in this specific case: when the initial guess is smaller, the 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



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 5 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

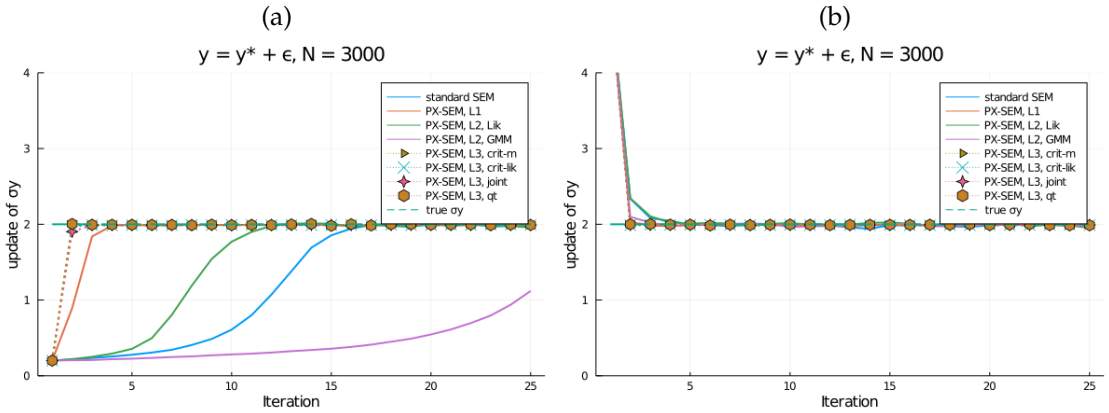
$$\hat{\sigma} = \widehat{\text{std}}\big(y - (y^* - k_1 y) * (pinv(y^* - \hat{k}_1 y) * y)\big)$$

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.

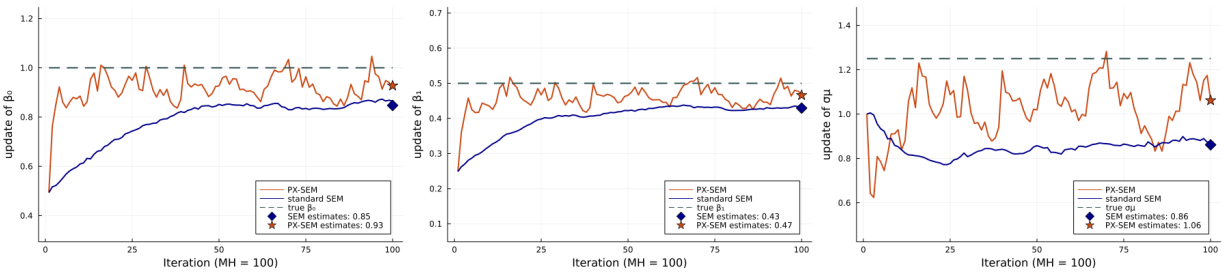
Comparison among all PX-SEM algorithms is shown in Figure B3. We can see that when the L model is more flexible (L3), at least in this case, even though the M-step estimates are not MLE, the performance is still good, and better than PX-SEM-L1. So there might be a trade-off between the flexibility of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

C Discrete Models with More Iterations

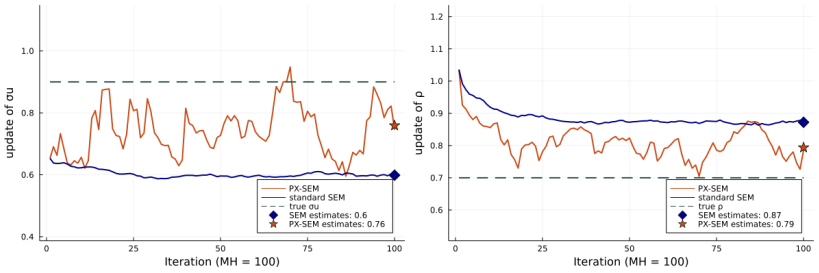
First I show only the 100 iterations. The result is very interesting. If we only conduct SEM algorithm for 100 iterations, it seems the changes along the iterations are rather small since 50th iteration such that one might thought it already converges. However, as we know the true value, we understand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

Figure C1: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses

(a) β_0 (b) β_1 (c) σ_μ



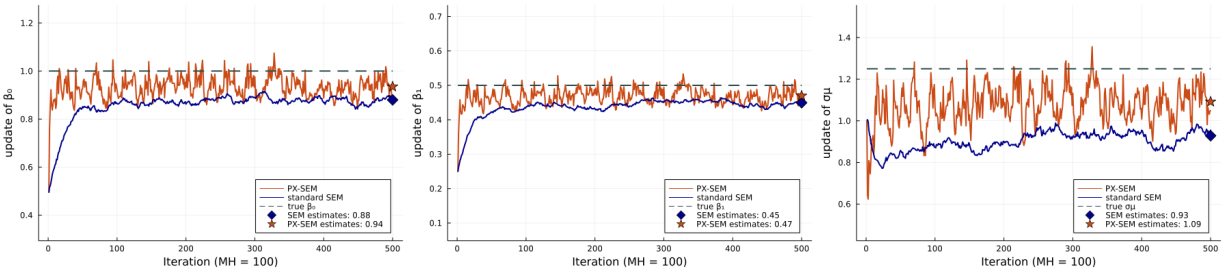
(d) σ_u (e) ρ



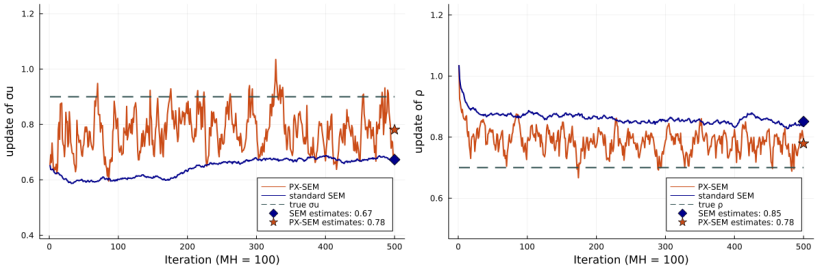
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 50 iterations. Based on informed initial guess.

Figure C2: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses, zoom to $[0, 500]$

(a) β_0 (b) β_1 (c) σ_μ

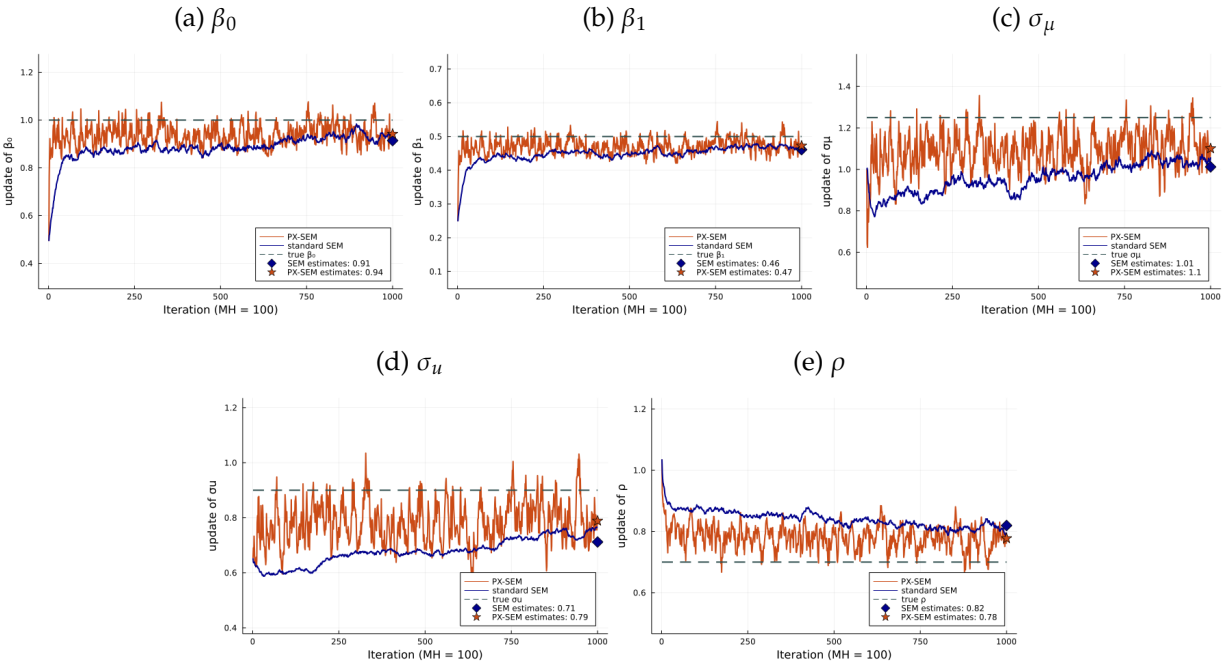


(d) σ_u (e) ρ



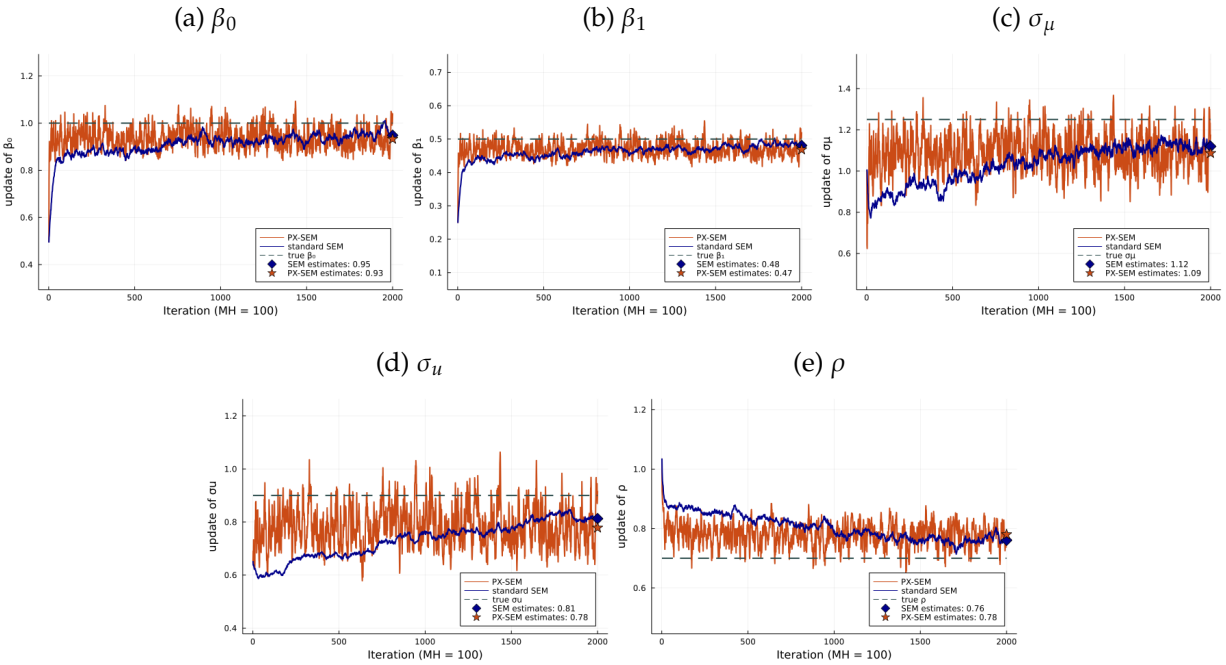
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 250 iterations. Based on informed initial guess.

Figure C3: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses, zoom to $[0, 1000]$



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.

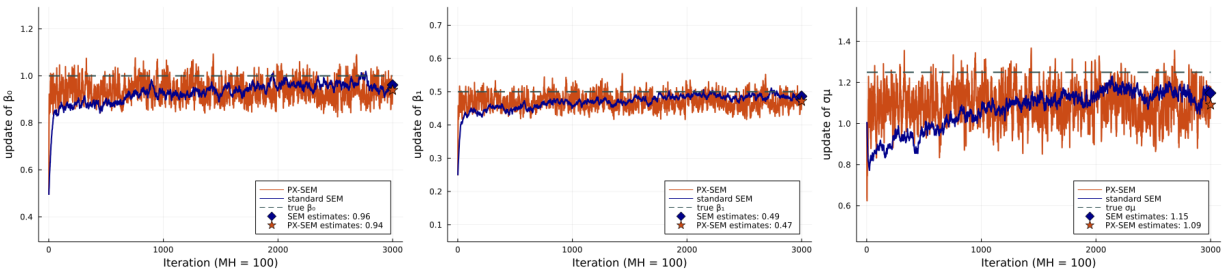
Figure C4: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses, zoom to $[0, 2000]$



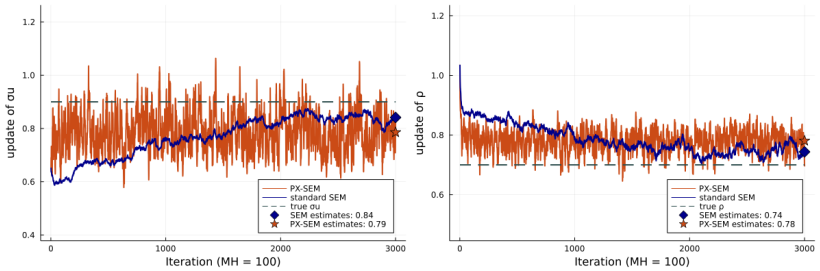
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 500 iterations. Based on informed initial guess.

Figure C5: 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



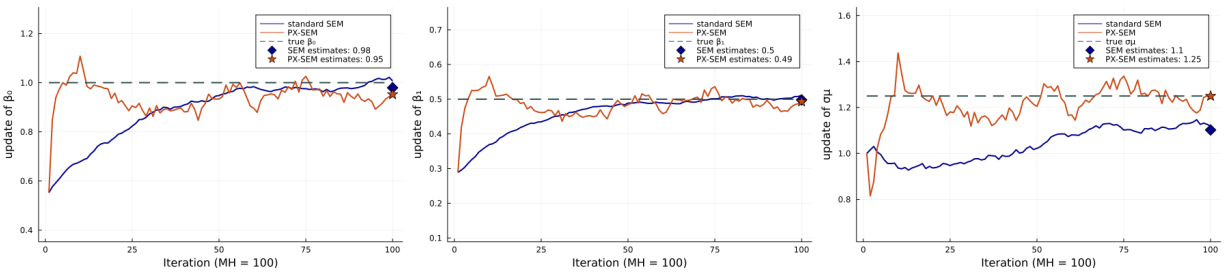
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 1000 iterations. Based on informed initial guess.

D More Simulations of Discrete Models

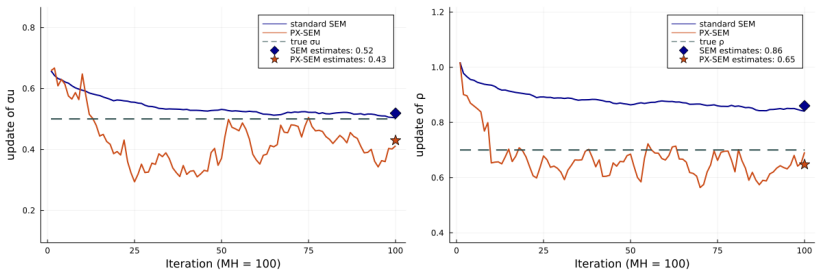
Informed Initial Guesses

Figure D1: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses

(a) β_0 (b) β_1 (c) σ_μ



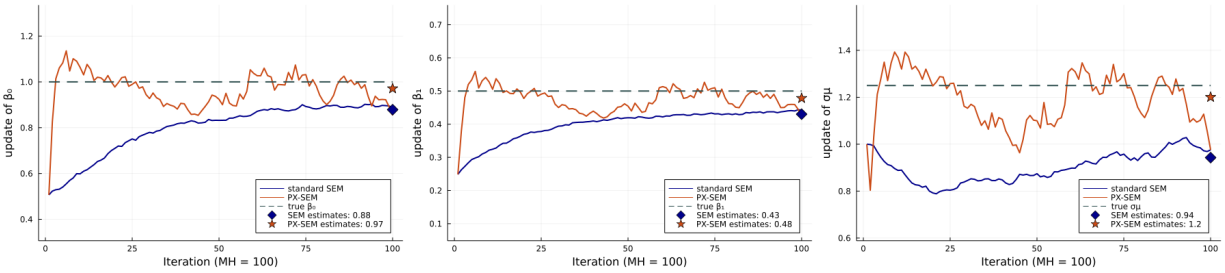
(d) σ_u (e) ρ



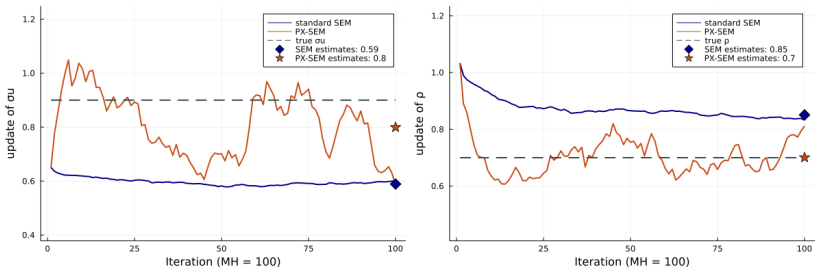
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 50 iterations. Based on informed initial guess.

Figure D2: SEM and PX-SEM iterations of $\Theta^{(s)}$ from informed guesses

(a) β_0 (b) β_1 (c) σ_μ



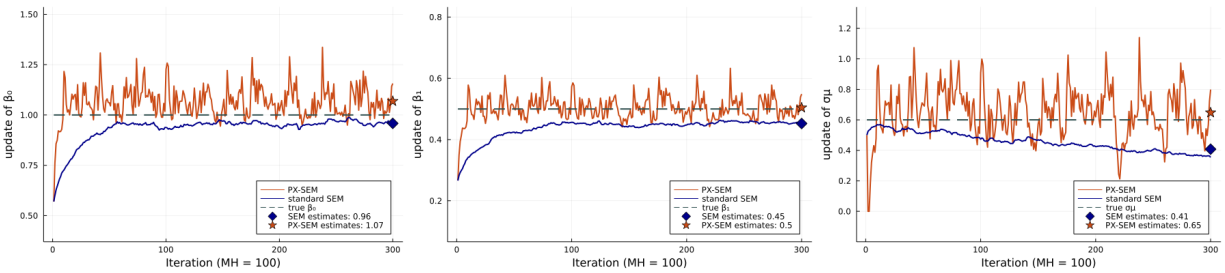
(d) σ_u (e) ρ



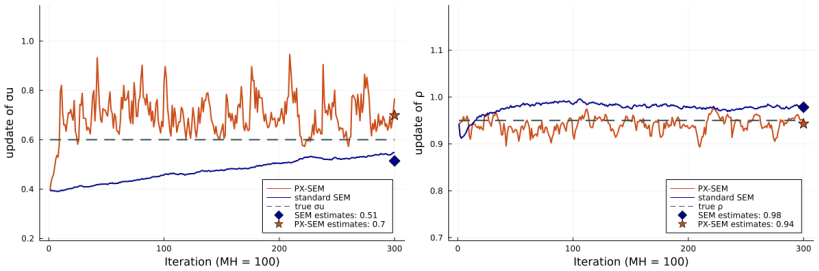
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 50 iterations. Based on informed initial guess.

Figure D3: 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

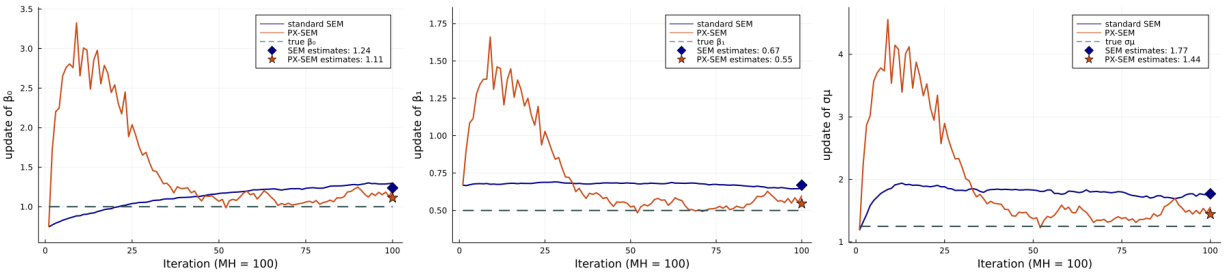


Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 300 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 150 iterations. Based on informed initial guess.

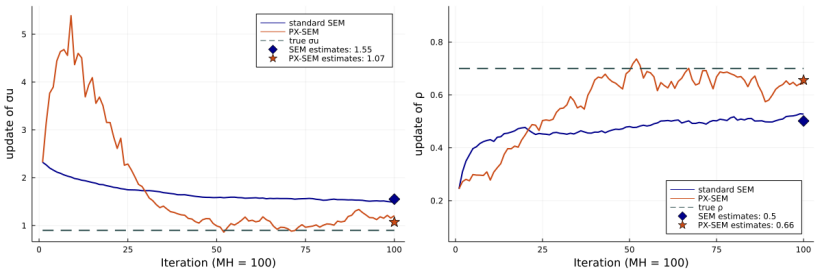
Random Initial Guesses

Figure D4: SEM and PX-SEM iterations of $\Theta^{(s)}$ from random initial guesses

(a) β_0 (b) β_1 (c) σ_μ

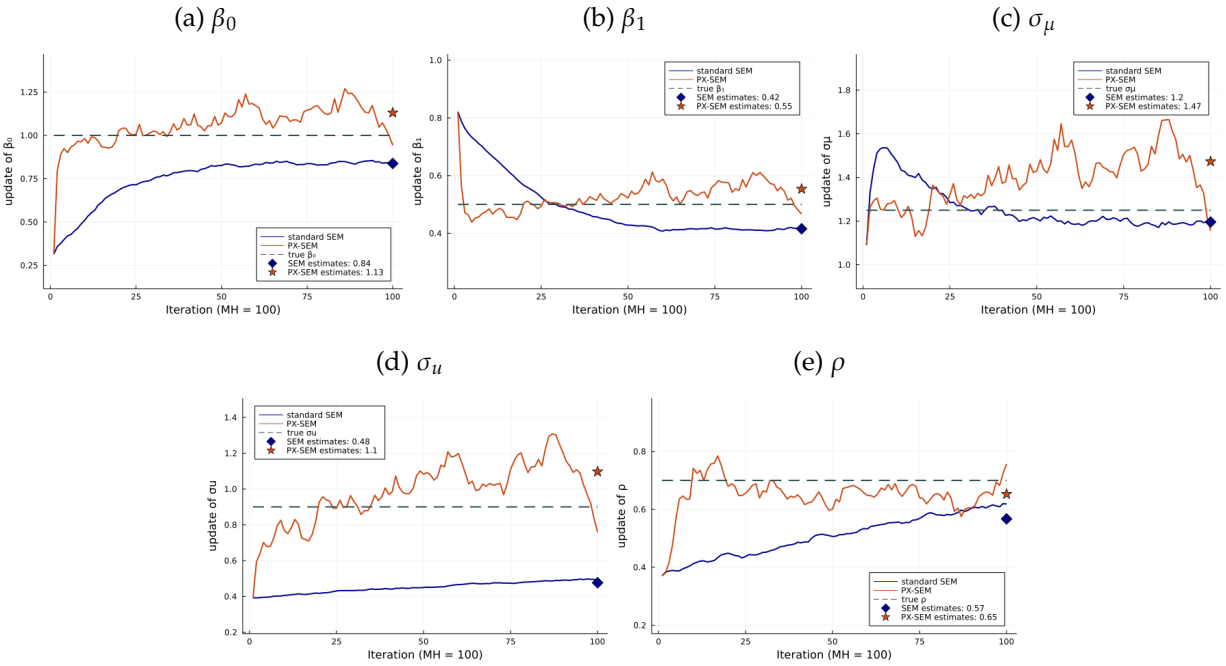


(d) σ_u (e) ρ



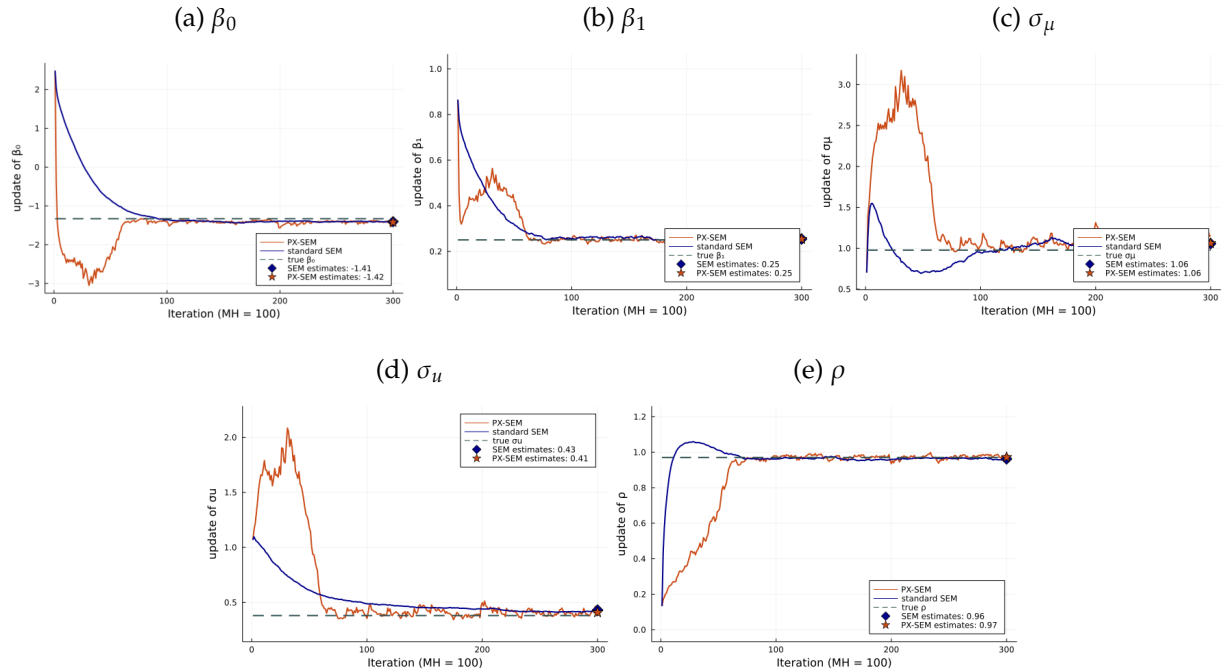
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 50 iterations. Random initial guess from lognormal distribution

Figure D5: SEM and PX-SEM iterations of $\Theta^{(s)}$ from random initial guesses



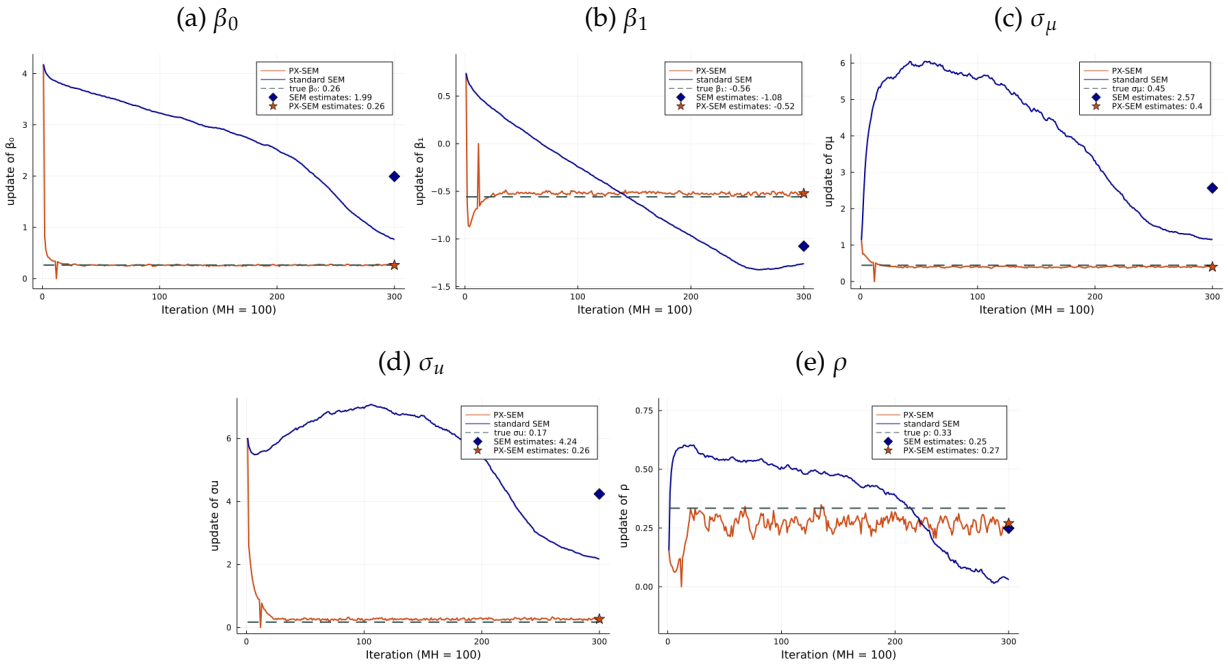
Notes: SEM (blue solid line) and PX-SEM (orange solid line) iterations based on 100 MH draws. True value are in green dash line. SEM estimates (blue diamond) and PXSEM estimates (orange star) are based on the average of last 50 iterations. Random initial guess from lognormal distribution

Figure D6: SEM and PX-SEM iterations of $\Theta^{(s)}$ from random initial guesses



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 D7: SEM and PX-SEM iterations of $\Theta^{(s)}$ from random initial guesses



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

E PX-M Step of Quantile Models Estimation

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 $\left[\hat{\mu}_i^* \quad \hat{v}_i^{*\prime} \quad \hat{\epsilon}_i^{*\prime}\right]' = \hat{\mathbf{A}}^{-1}\left[\mu_i \quad v_i' \quad \epsilon_i'\right]'$ 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 $y_{it} - \hat{\mu}_i^*$

   First **redefine** $v_{it} = y_{it} - \hat{\mu}_i^* - \epsilon_{it}$, and then complete the following loop:

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

   Define

   $$res_{ik}^{\epsilon} \equiv \epsilon_{ik} - Z_i'\left(\sum_i Z_i Z_i'\right)^{-1}\left(\sum_i Z_i \epsilon_{ik}\right)$$

   and

   $$res_{ik} \equiv y_{ik} - \hat{\mu}_i^* - Z_i'\left(\sum_i Z_i Z_i'\right)^{-1}\left(\sum_i Z_i\left(y_{ik} - \hat{\mu}_i^*\right)\right)$$

   where $Z_i \equiv [\hat{v}_{i1}^*, ..., \hat{v}_{i,k-1}^*, v_{i,k+1}, ..., v_{i,T}, \hat{\epsilon}_{i1}^*, ..., \hat{\epsilon}_{i,k-1}^*, \epsilon_{i,k+1}, ..., \epsilon_{i,T}]'$

   Then we define

   $$\delta_{ik} \equiv res_{ik}^{\epsilon} - \mathbf{a}^{\epsilon} res_{ik}$$

   Finally, we 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 $S_i = [Z_i', \delta_{ik}]'$ which is a function 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 $[\hat{v}_{i1}^*, ..., \hat{v}_{i,k-1}^*, v_{i,k+1}, ..., v_{i,T}, \hat{\epsilon}_{i1}^*, ..., \hat{\epsilon}_{i,k-1}^*, \epsilon_{i,k+1}, ..., \epsilon_{i,T}]'$ and δ_{ik} are 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 $\hat{\mathbf{a}}^{\mu}, \hat{\mathbf{a}}^{\epsilon} = \arg\min_{a^{\mu}, a^{\epsilon}} G\big(\hat{\mu}_i^*(a^{\mu}, a^{\epsilon}), \hat{v}_i^*(a^{\mu}, a^{\epsilon}), \hat{\epsilon}_i^*(a^{\mu}, a^{\epsilon})\big)$.

Function $G(\cdot)$ describes who 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 $t \in [3, ..., T]$,

$$coef_t = \left(\sum_i res_i^r res_i^{r\prime}\right)^{-1}\left(\sum_i res_i^r res_{it}^v\right)$$

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[1]

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 $\Rightarrow$ τ_{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.

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

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

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

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

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

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

F 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 $\hat{\mu}_i^*$, $\hat{v}_i^*$, and $\hat{\epsilon}_i^*$ from the E-step draws by imposing zero correlation, and then conduct quantile regressions based on $\hat{\mu}_i^*$, $\hat{v}_i^*$, 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 $\hat{\mu}_i^*$, $\hat{v}_i^*$, 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 $[\hat{\mu}_i^* \;\; \hat{v}_i^* \;\; \hat{\epsilon}_i^*] = \mathbf{A}^{-1}[\mu_i \;\; v_i \;\; \epsilon_i]$ by imposing orthogonality

1. replace $\mu = \mu - X(pinv(x)\mu)$, $v_t = v_t + X(pinv(x)\mu)$ 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 $v_t = y_t - \hat{\mu}^* - \epsilon_t$. 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 $x = 0.5 \pm \sqrt{\mathbf{a}^{\mu}(1 - \mathbf{a}^{\mu})\,\mathrm{var}(\mu)/\mathrm{var}(\frac{1}{T}\sum_t(y_t - \mu)) + 0.25}$ 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 $\hat{v}_t^*$. Iterate over t:

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

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

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

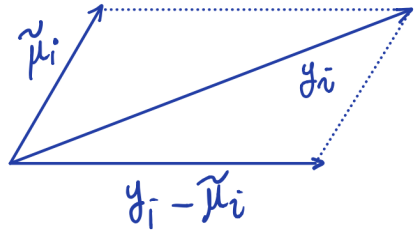
(c) Adjust the direction&length: gen $\hat{\epsilon}_k^* = \mathbf{a}^\epsilon res_k^\epsilon + x res_k^v$. Given $\mathbf{a}^\epsilon$ and orthogonality between res_k^ϵ and res_k^v, x can be solved (in step 2).

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

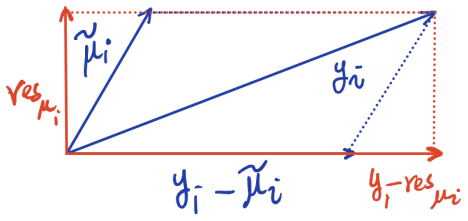
The procedure can be explained using the following figures. Assume there are only two variables. Figure F1a shows the draws from the E-step, where two variables are not orthogonal. Steps 3a and 3b aim to orthogonalize two variables and achieve the result as in Figure F1b. Then the auxiliary parameter is the length in current direction, and the finally vector is decided by imposing orthogonality as indicated in Figure F1c. 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 F1: Procedures explained in figures

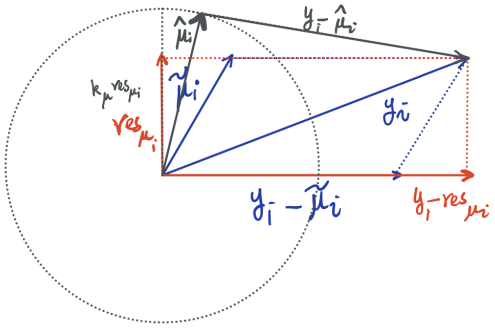
(a) original



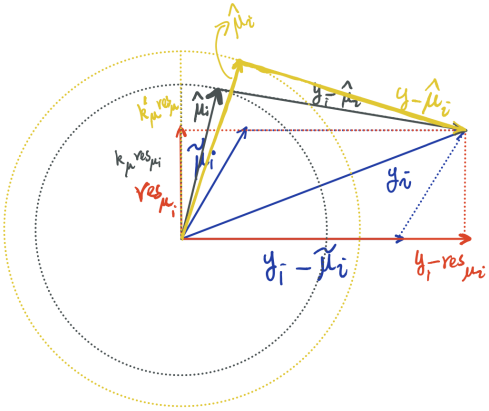
(b) orthogonalization



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



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

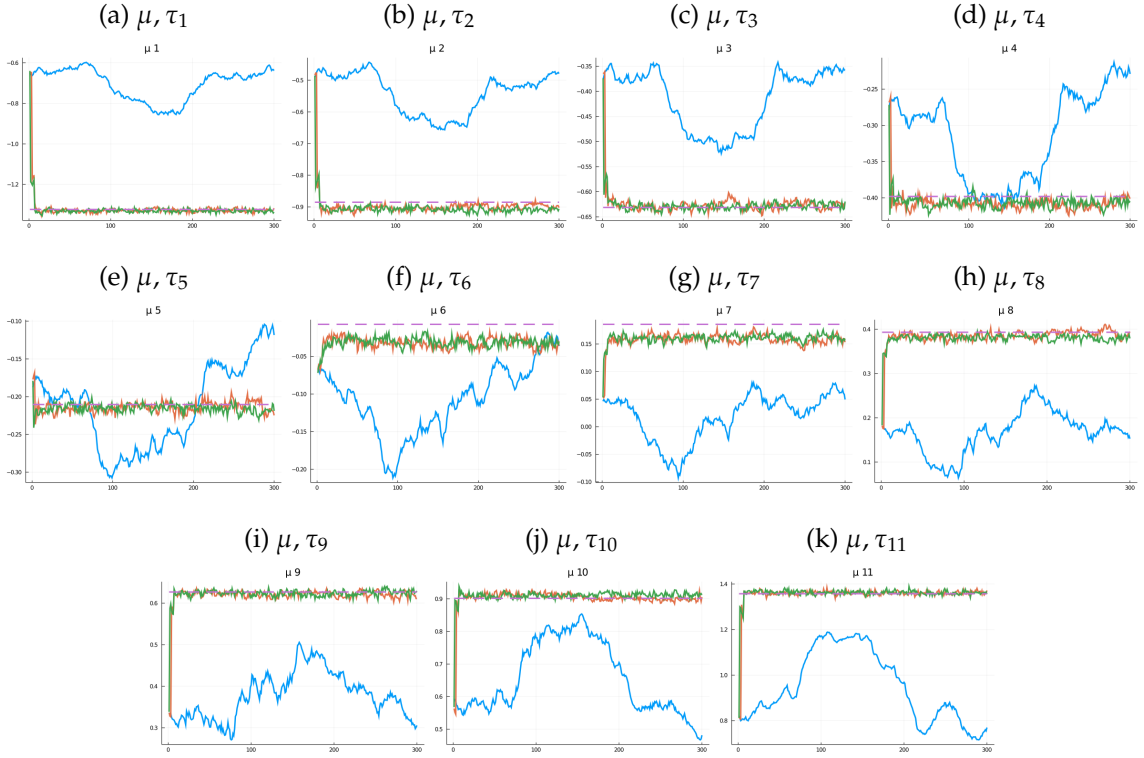


G Preliminary Results for Quantile Models

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 $v_{it} = Q(v_{i,t-1}, v^2_{i,t-1}, u_{it})$ allows for nonlinear conditional mean and nonlinear persistence.

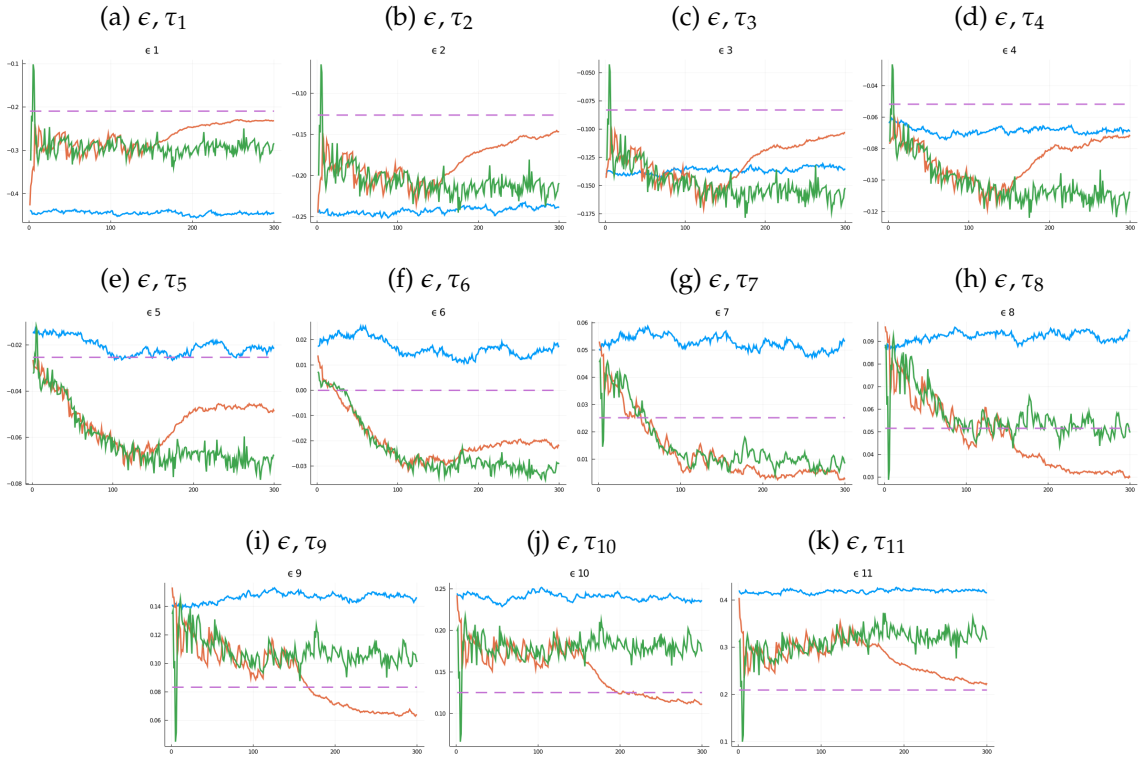
From the results shown below, we can see that SEM does not converge within 300 iterations, whereas both PX-SEM moves toward the region near the true value quickly. Based on the results of first 300 iterations, we plot the estimated distribution of each component. Not surprisingly, SEM does not capture 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 v_{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 G1: SEM, PX-SEM, PX-SEM+SEM iterations of distribution parameters of μ_i



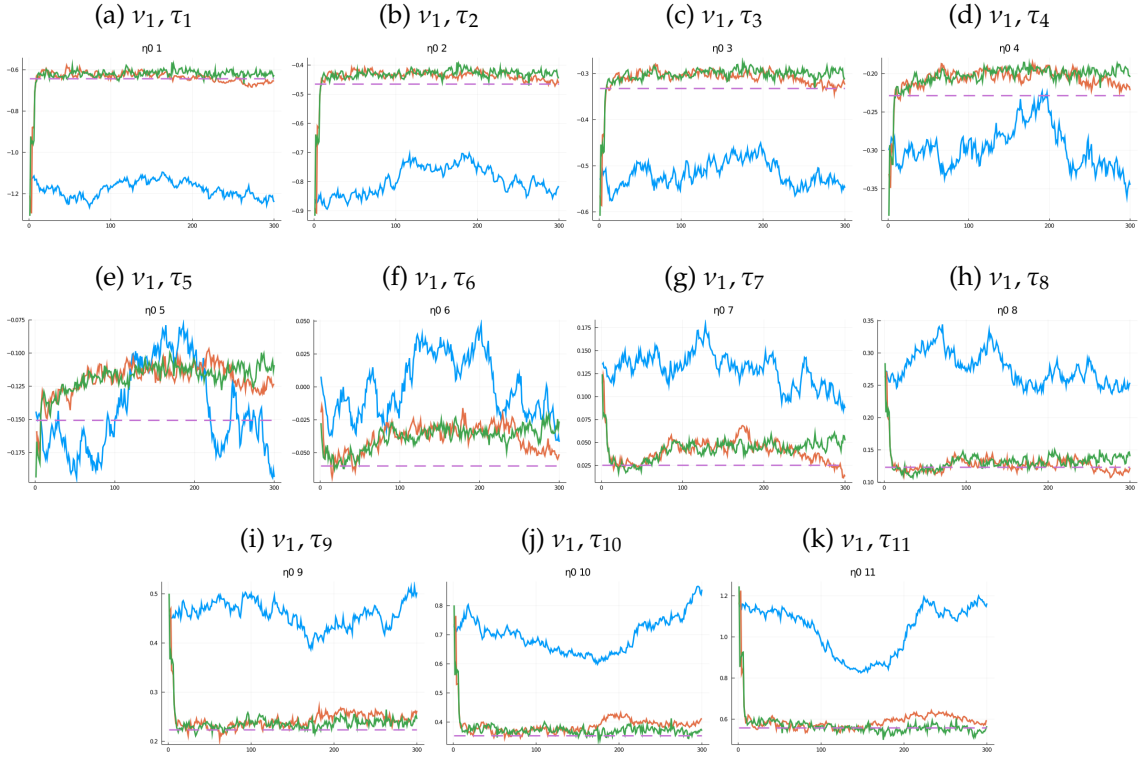
Notes: SEM—blue solid, PX-SEM — green solid, PX-SEM+SEM —orange solid, true values—pink dash.

Figure G2: SEM, PX-SEM, PX-SEM+SEM iterations of distribution parameters of ϵ_i



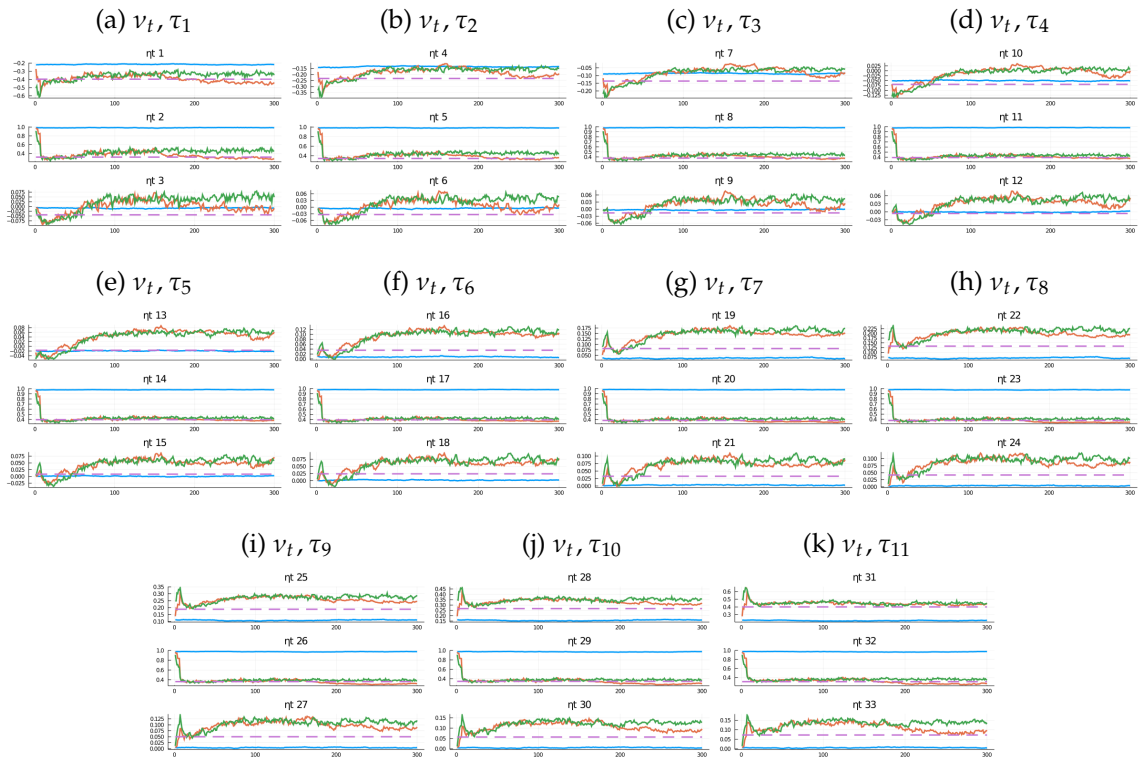
Notes: SEM—blue solid, PX-SEM — green solid, PX-SEM+SEM —orange solid, true values—pink dash.

Figure G3: 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 G4: 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 G5: 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

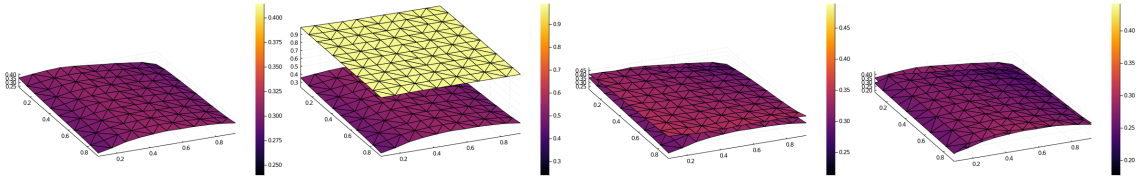
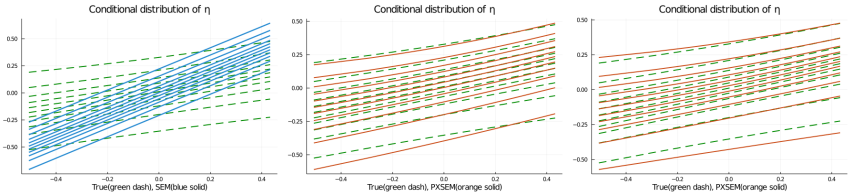


Figure G6: SEM, PX-SEM, PX-SEM+SEM estimates of conditional distribution of $\nu_t|\nu_{t-1}, \ldots$

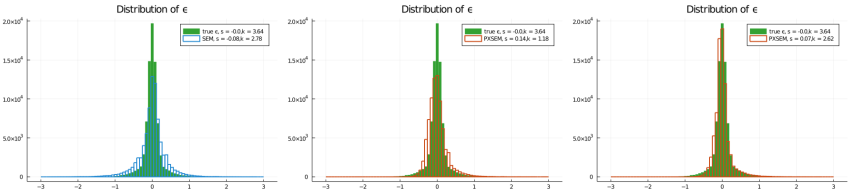
(a) ν sem (b) ν pxsem (c) ν pxsem-sem



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

Figure G7: SEM, PX-SEM, PX-SEM+SEM estimates of distribution of ϵ_i

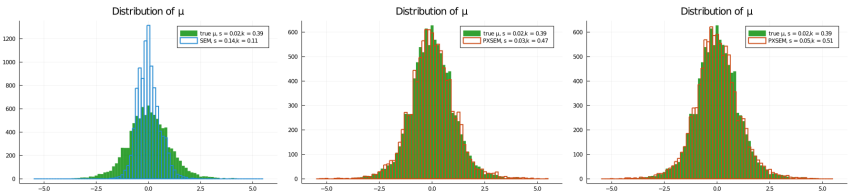
(a) ϵ sem (b) ϵ pxsem (c) ϵ pxsem-sem



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

Figure G8: 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