Taylor & Francis
Taylor & Francis Group

Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ubes20

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

Siqi Wei

To cite this article: Siqi Wei (15 Jul 2024): Estimating Latent-Variable Panel Data Models Using Parameter-Expanded SEM Methods, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2024.2365783

To link to this article: https://doi.org/10.1080/07350015.2024.2365783

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2024, VOL. 00, NO. 0, 1–14
https://doi.org/10.1080/07350015.2024.2365783

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

Siqi Wei

IE University, Madrid, Spain

ABSTRACT

This article presents new estimation algorithms for three types of dynamic panel data models with latent variables: factor models, discrete choice models, and persistent-transitory quantile processes. The new methods combine the parameter expansion (PX) ideas of Liu, Rubin, and Wu with the stochastic expectation-maximization (SEM) algorithm in likelihood and moment-based contexts. The goal is to facilitate convergence in models with a large space of latent variables by improving algorithmic efficiency. This is achieved by specifying expanded models within the M step. Effectively, we are proposing new estimators for the pseudo-data within iterations that take into account the fact that the model of interest is misspecified for draws based on parameter values far from the truth. We establish the asymptotic equivalence of the likelihood-based PX-SEM to an alternative SEM algorithm with a smaller expected fraction of missing information compared to the standard SEM based on the original model, implying a faster global convergence rate. Finally, in simulations we show that the new algorithms significantly improve the convergence speed relative to standard SEM algorithms, sometimes dramatically so, by reducing the total computing time from hours to a few minutes.

ARTICLE HISTORY
Received October 2023
Accepted May 2024

KEYWORDS
Algorithmic efficiency;
Discrete choice model;
Dynamic factor model;
Dynamic quantile model;
PX-EM; Stochastic EM

1. Introduction

This article presents new estimation algorithms for dynamic panel data models with latent variables. Dynamic panel data models are widely used in applied work today. They tend to exhibit many latent variables over multiple periods (e.g., time-invariant, persistent, and transitory components), which are important to capture unobserved heterogeneity and dynamic responses (Arellano and Bonhomme 2017). However, the presence of latent variables brings challenges to the estimation.

Iterative methods like the stochastic expectation-maximization (SEM) algorithm can be useful tools for estimating models with latent variables (Diebolt and Celeux 1993).[1] Specifically, as a simulated version of the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977), SEM iterates through an E-step where we draw latent variables from the posterior distribution of the model of interests given observables, and an M-step where we estimate the model as if the draws were observables until the parameters *converge* to the stationary distribution. It simplifies the estimation as it replaces the complex optimization problem, which involves multiple integrals due to latent variables, with a series of much simpler optimization problems under pseudo-complete data.

However, the slow convergence, an often voiced criticism of EM and its variants, tends to diminish its practical appeal. Indeed, the slow convergence issue in practice is even more pronounced: researchers often need to run the algorithms multiple times with different initial guesses and select the result based on criteria such as the likelihood value, to mitigate the negative effects of a "bad" initial guess and to address the possibility of the algorithm converging to a local maximum. Recent research has explored alternative samplers for latent variables when performing the E-step to improve sampling efficiency and stability.[2] In contrast, this article focuses on the potential improvement in the M-step.

In this article, we develop a new estimation method, the PX-SEM algorithm, by combining the parameter expansion ideas in Liu, Rubin, and Wu (1998) with the SEM algorithm. The goal is to facilitate convergence in models with a large space of latent variables by improving algorithmic efficiency. Even though the general concept of the PX-SEM algorithm applies to various models, we focus on three types of dynamic panel data models containing rich latent variable structures, where slow convergence issues are exacerbated, with the expectation that it can be particularly fruitful: (a) dynamic factor models, (b) random effects discrete choice models with persistent and transitory components, and (c) persistent-transitory dynamic quantile models with individual effects.[3]

[1] Arellano and Bonhomme (2017) discusses the potentials of SEM in nonlinear panel data analysis.

[2] For instance, Arellano et al. (2023) develops a Sequential Monte Carlo sampler for the E step.

[3] Wei (2022) applies the algorithms developed in this article to a substantive analysis of the earnings and employment dynamics of older workers, which

CONTACT Siqi Wei *siqi.wei@ie.edu* Calle de Maria de Molina, 12, 28006, Madrid, Spain

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/UBES*.

The PX-SEM algorithm consists of two steps: an E-step, where we draw values of latent variables from the posterior distribution, and a PX-M-step, where we update parameters. Having the same E-step as the SEM algorithm, PX-SEM replaces the SEM M-step estimator with a more robust one, taking into account the possibility that E-step draws could violate model assumptions when parameter guesses are far from the true value. The PX-M step estimator is able to leverage additional information from the model itself, effectively "correcting" the M step updates in progressing to more accurate ones.

To implement the PX-SEM algorithm, one must construct an expanded model, the *L model*, which needs to satisfy two conditions. First, the L model must nest the original model, the *O model*. Second, there must exist a reduction function, a mapping from the L model parameters to the O model parameters, keeping the observed-data likelihood unchanged. After constructing a suitable L model, we can iterate between the E step and the PX-M step, which involves (a) estimating the L model and (b) mapping back to the O model parameters through the reduction function.

There are different ways to construct L models. All else being equal, a more flexible L model should improve the convergence rate. However, since our ultimate goal is to reduce the total computing time, we also need to consider the time spent in each iteration for estimating the L model and converting it to the O model. Therefore, taking these two factors into account, this article proposes a method to expand the model *linearly*. Linear expansion addresses the potential violation of zero-correlation assumptions.

In terms of statistical properties, Liu, Rubin, and Wu (1998) proves the monotone convergence of the parameter-expanded EM algorithm and its superior rate of convergence relative to its parent EM. By combining the results of Nielsen (2000) and Arellano and Bonhomme (2016), this article establishes the asymptotic equivalence of the likelihood-based PX-SEM to an alternative SEM algorithm with a smaller expected fraction of missing information compared to the standard O model based SEM, implying a faster global convergence rate and a smaller variance for the limiting stationary distribution. Finally, in the simulations, we show that PX-SEM can significantly improve algorithmic efficiency compared to the standard SEM algorithm, sometimes dramatically so. For example, in our numerical calculations for discrete choice and quantile models, SEM has still not converged even after running for 50–80 min whereas PX-SEM converges within 2–3 min.

This article belongs to an expanding literature that considers the application of the EM algorithm (Dempster, Laird, and Rubin 1977) and its variants in estimating latent variable models (Diebolt and Celeux 1993; Liu, Rubin, and Wu 1998; Arcidiacono and Jones 2003; Pastorello, Patilea, and Renault 2003; Arellano and Bonhomme 2016; Chen 2016; Arellano et al. 2023, among others). This article contributes to this literature in two ways. First, by developing a new estimation method, PX-SEM, which combines the parameter expansion idea with the SEM algorithm.[4] The method offers appealing theoretical properties and the potential to enhance algorithmic efficiency, which is particularly valuable for complex models such as nonlinear panel data models, where SEM may encounter slow convergence issues. Second, the article proposes a specific class of *linear* expansions for implementing PX-SEM and develops new estimation algorithms for three types of latent-variable panel data models with enhanced algorithmic efficiency.

The article proceeds as follows. Section 2 illustrates the difference between the standard stochastic EM algorithm and the parameter-expanded stochastic EM algorithm using a simple toy model. In Section 3, a formal definition of PX-SEM is provided, along with a discussion of its statistical properties and implementation based on linear expansions. Sections 4–6 develop PX-SEM methods for three types of latent-variable panel data models: dynamic factor models, discrete choice models, and persistent-transitory dynamic quantile models, respectively. Finally, Section 7 concludes.

2. Toy Model

Based on a simple toy model, this section compares the parameter-expanded stochastic EM (PX-SEM) algorithm with the standard stochastic EM (SEM) algorithm and provides intuitions behind PX-SEM.

Consider the following model we want to estimate, denoted as the O model:

$$y_i = y_i^* + \epsilon_i, \quad \text{where } \begin{bmatrix} y_i^* \\ \epsilon_i \end{bmatrix} \overset{\text{iid}}{\sim} N\left(0, \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix}\right). \qquad (O\ Model)$$

The observed outcomes are $y_1, \ldots, y_N$, and the latent variables whose distribution is of interest are $y_1^*, \ldots, y_N^*$. The only unknown parameter is the standard deviation σ.

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

1. Stochastic E step: Draw y_i^* from the posterior distribution $f_O(y_i^* | y_i; \hat{\sigma}^{(s)})$
2. M step: Estimate the O model and update $\hat{\sigma}^{(s+1)}$, that is $\hat{\sigma}^{(s+1)} = \widehat{\text{std}}(y_i^*)$

where $f_O(\cdot)$ is the density function of O model. The final estimator is the average of the last S^0 iterations $\hat{\sigma} = \frac{1}{S^0} \sum_{S-S^0+1}^{S} \hat{\sigma}^{(s)}$.

The nonstochastic version, the EM algorithm, is effective because it improves the observed-data likelihood in each iteration:

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

where $Q(\hat{\sigma}^{(s+1)} | \hat{\sigma}^{(s)}) = \sum_i \int \log f_O(y_i, y_i^*; \hat{\sigma}^{(s+1)}) f_O(y_i^* | y_i; \hat{\sigma}^{(s)}) dy_i^*$.[5]

brings together elements of the three types of panel models considered here.

[4] Liu, Rubin, and Wu (1998) is based on the EM algorithm; Liu and Wu (1999) applies the parameter expansion technique to Bayesian inference; Lavielle and Meza (2007) combines the parameter expansion technique with Monte Carlo EM (Wei and Tanner 1990).

[5] See Wu (1983) for more discussions.

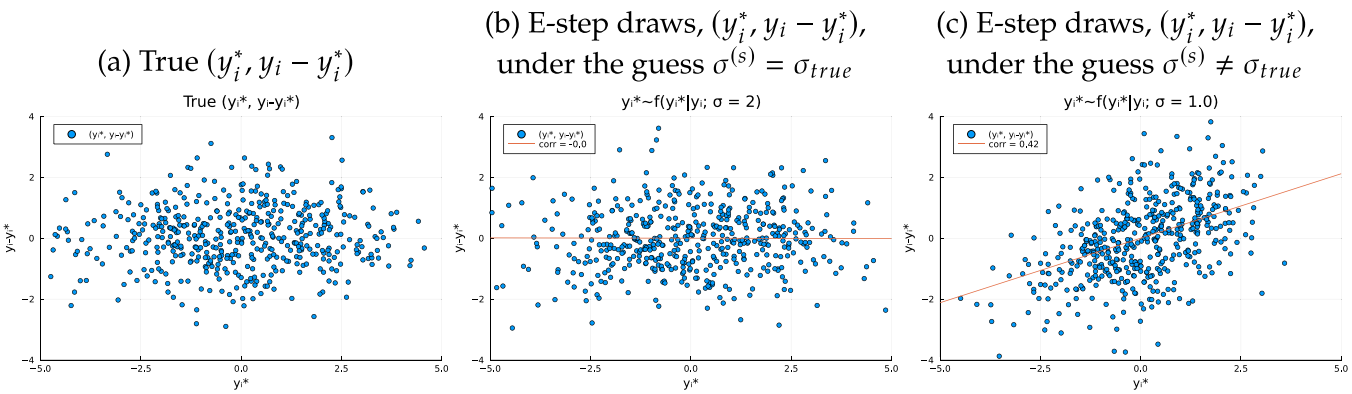

Figure 1. Data and E-step draws under different guess values of σ.

PX-SEM. Now we introduce the PX-SEM algorithm. Like SEM, PX-SEM comprises an E step for draws of latent variables and an M step for updating parameters. While sharing the same E-step as SEM, PX-SEM's M step involves (a) estimating an expanded model, the L model, and (b) mapping L model parameters to the O model parameters.

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

$$y_i = y_i^* + \epsilon_i, \qquad (L\ Model)$$

$$\text{where } \begin{bmatrix} y_i^* \\ \epsilon_i \end{bmatrix} \overset{\text{iid}}{\sim} N\left(0, K \begin{bmatrix} \sigma^2 & 0 \\ 0 & 1 \end{bmatrix} K'\right),$$

$$K = \begin{bmatrix} k & 0 \\ 1-k & 1 \end{bmatrix}.$$

In addition to σ, the L model also contains an auxiliary parameter k. It is easy to verify 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 $k \neq 1$, L model expands the O model by allowing for a nonzero correlation between y_i^* and ϵ_i, as $\text{cov}(y_i^*, \epsilon_i) = k(1-k)\sigma^2$. Moreover, the L model is unidentified with observables: for any L model with parameter values σ_L and k_L, the O model with parameter value $\sigma = \sigma_L$ has the same observed data likelihood, that is $f_O(y_i; \sigma_L) = f_L(y_i; \sigma_L, k_L)$.[6]

To implement PX-SEM, we begin with an initial guess of the unknown parameter $\hat{\sigma}^{(0)}$, and then iterate the following two steps for $s = 0, 1, \ldots, S$ until the convergence of $\hat{\sigma}^{(s)}$ to the stationary distribution:

1. Stochastic E step: Draw y_i^* from the posterior distribution $f_O(y_i^* | y_i; \hat{\sigma}^{(s)})$
2. PX-M step: Update parameters by
   (a) Estimate the L model: $\hat{k}_L = \frac{\widehat{\text{var}}(y_i^*)}{\widehat{\text{cov}}(y_i^*, y_i)}, \hat{\sigma}_L^{(s+1)} = \frac{\widehat{\text{std}}(y_i^*)}{|\hat{k}_L|}$
   (b) Reduction: mapping from $(\hat{\sigma}_L^{(s+1)}, \hat{k}_L)$ to $\hat{\sigma}^{(s+1)}$ maintaining the observed-data likelihood, that is $f_O(y_i; \hat{\sigma}^{(s+1)}) = f_L(y_i; \hat{\sigma}_L^{(s+1)}, \hat{k}_L)$, and thus $\hat{\sigma}^{(s+1)} = \hat{\sigma}_L^{(s+1)}$

The final estimator is the average of the last S^0 iterations: $\hat{\sigma} = \frac{1}{S^0} \sum_{S-S^0+1}^{S} \hat{\sigma}^{(s)}$. As we see, the difference between the two methods lies in the estimators of the M step: the PX-SEM estimator is adjusted by $\frac{1}{|\hat{k}_L|}$.

[6]Since k does not affect the L model observed data likelihood, $f_O(y_i; \sigma_L) = f_L(y_i; \sigma_L, 1) = f_L(y_i; \sigma_L, k_L)$

Figure 1 illustrates how PX-SEM has the potential to enhance algorithmic efficiency through the utilization of the auxiliary parameter k. Specifically, Figure 1(a) depicts a scatterplot of simulations generated by the data generating process (DGP), the O model with a true value of $\sigma = 2$. The X-axis and Y-axis display y_i^* and ϵ_i, respectively, but only their sum $y_i = y_i^* + \epsilon_i$ is used for estimation.

Figure 1(b) is the scatterplot of $(y_i^*, y_i - y_i^*)$ where y_i^* is the E-step draws under the true value, that is $y_i^* \sim f_O(y_i^* | y_i; \sigma = 2)$, and Figure 1(c) is the scatterplot of $(y_i^*, y_i - y_i^*)$ where y_i^* is the E-step draws under an incorrect value $\sigma = 1$, that is $y_i^* \sim f_O(y_i^* | y_i; \sigma = 1)$. Contrary to the case of Figure 1(b), where E-step draws are generated under correct guess, and there is no significant correlation between y_i^* and $y_i - y_i^*$, Figure 1(c) presents a significant positive correlation between E-step draws y_i^* and $y_i - y_i^*$, which is assumed to be zero by the O model. This "false" positive correlation arises because the draws are taken under a wrong condition that the variance of y_i^* should not deviate significantly from one.

As a consequence, for the E-step draws in Figure 1(b), both M-step and PX-M step estimators are consistent: the SEM one is under a correct constraint $k = 1$. However, in the case of Figure 1(c), the M-step of SEM ignores the violation of the zero-correlation assumption at the current draws, while PX-SEM takes into account the "false" correlation by adding the parameter k. By construction, the extra flexibility induced by k leads to a better model fit of the PX-M step for the current draws and thus, a larger pseudo complete data likelihood. As we will show in Section 3, similar to EM, the PX-EM can improve observed-data likelihood in each iteration by transmitting gains from pseudo-complete data likelihood. Therefore, improving model fit as in the PX-M step essentially equates to improving the lower bound of the inequality in (1). Finally, by mapping the L model parameters to the O model parameters keeping the observed-data likelihood unchanged, we preserve the "gains" in likelihood. Intuitively, PX-SEM replaces the SEM M-step with a more robust estimator that leverages additional model information, namely, there is a linear correlation between the E-step draws y_i^* and $y_i - y_i^*$, even though there should not be. Appendix A provides more explanation with illustrative figures.

Regarding the choice of the L model, there are many different ways of expanding the O model. Flexible expansion should help increase the convergence rate. Appendix A compares different L

models and PX-M step estimators using the toy model. However, in practice, when we estimate more complex models, we must consider how easily we can estimate the L model and reduce it to the O model to avoid spending too much time in each iteration and increasing the total computing time. With this consideration in mind, we propose a linear expansion method in Session 3 and discuss its applications in Sessions 4–6.

As a final comment, the PX-SEM algorithm aims to enhance algorithm efficiency, even when E-step draws are appropriately obtained. For instance, as demonstrated in Appendix A, PX-SEM improves the convergence rate when the E-step draws are based on direct sampling. In more complex models where direct sampling is not feasible, and methods such as MCMC are required, the PX-SEM algorithm demonstrates lower sensitivity to poorly generated E-step draws. This can be justified if one of the ways that "bad" draws manifest themselves is through violations of model assumptions. Moreover, an MCMC-based E-step generally requires more time for each iteration, so reducing the iterations needed for convergence can significantly reduce the total computing time. For example, as shown in Sections 5 and 6, while SEM can take more than 3000 sec without showing signs of convergence, PX-SEM converges within a couple of minutes.

3. Parameter Expanded Stochastic EM Algorithm

The section begins by defining the PX-SEM algorithm. Following that, we discuss its statistical properties and explore the reasons behind its potential to enhance algorithmic efficiency. Specifically, we establish the asymptotic equivalence between PX-SEM and an alternative SEM with a reduced expected fraction of missing information. Finally, we propose a general approach, the linear expansion method, for the implementation.

3.1. Definition of PX-SEM Algorithm

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

$$E\Big(\int \Psi_O(Y_i^*, W_i;\overline{\theta}) f_O(Y_i^*|W_i;\overline{\theta})dY_i^*\Big) = 0. \quad (2)$$

Define $\hat{\theta}$ as the solution of the integrated moment restrictions of the original O model, $\sum_i \Big(\int \Psi_O(Y_i^*, W_i;\hat{\theta}) f_O(Y_i^*|W_i;\hat{\theta})dY_i^*\Big) = 0$, a sample analogy to (2). In the case of a likelihood-based algorithm, we know that $\hat{\theta}$ is the MLE.

Denote the expanded model, the L model, as $f_L(Y_i|X_i;\theta, K) = \int_{Y_i^*} f_L(Y_i, Y_i^*|X_i;\theta, K)dY_i^*$, where K represents the auxiliary parameter vector. The expanded L model needs to satisfy two conditions: (a) The L model nests the O model: There exists $K = K_0$ such that $f_O(Y_i, Y_i^*|X_i;\theta) = f_L(Y_i, Y_i^*|X_i;\theta, K_0)$, $\forall\theta$, and (b) Existence of reduction function: There exists a mapping from the L model parameters to the O model parameters, the reduction function $\theta = R(\theta_L, K)$, such that the observed-data likelihood is preserved: $f_O(Y_i|X_i;R(\theta_L, K)) = f_L(Y_i|X_i;\theta_L, K)$, $\forall\theta_L, K$.

Let $\Psi_L^\theta(\cdot)$ denote the score function of the L model with respect to θ in the case of the likelihood-based PX-SEM algorithm and the same moment restrictions as $\Psi_O(\cdot)$ in the case of the moment-based one. Under condition (a), we have $\Psi_L^\theta(Y_i^*, W_i;\theta, K_0) = \Psi_O(Y_i^*, W_i;\theta)$, and thus $E(\Psi_L^\theta(Y_i^*, W_i;\overline{\theta}, K_0)) = 0$. Additionally, assuming that K is identified when we observe Y_i^*, meaning that there exists $\Psi_L^K(\cdot)$ such that $E(\Psi_L^K(Y_i^*, W_i;\overline{\theta}, K_0)) = 0$, we then have $E(\Psi_L(Y_i^*, W_i;\overline{\theta}, K_0)) = 0$, where $\Psi_L(\cdot) = [\Psi_L^\theta(\cdot)' \ \Psi_L^K(\cdot)']'$. By the law of iterated expectations, this implies:[7]

$$E\Big(\int \Psi_L(Y_i^*, W_i;\overline{\theta}, K_0) f_O(Y_i^*|W_i;R(\overline{\theta}, K_0))dY_i^*\Big) = 0. \quad (3)$$

Definition of the PX-SEM algorithm. Before we outline the general steps of PX-SEM, let us take a look at the SEM algorithm for comparison. SEM is an iterative algorithm where, in the E step, we make draws of latent variables Y_i^* from the posterior distribution $f_O(Y_i^*|W_i;\hat{\theta}^{(s)})$ under the parameter guess $\hat{\theta}^{(s)}$, and in the M step, we update it to $\hat{\theta}^{(s+1)}$, which satisfies $\sum_i \Psi_O(Y_i^*, W_i;\hat{\theta}^{(s+1)}) = 0$. This stochastic version differs from the original EM algorithm as it replaces the integral in (2) with latent draws.

In contrast, the PX-SEM algorithm proposes iterations that are better linked to (3): while we still make draws of latent variables Y_i^* from the posterior distribution $f_O(Y_i^*|W_i;\hat{\theta}^{(s)})$ under the parameter guess $\hat{\theta}^{(s)}$, we use the expanded model to update the parameter to $\hat{\theta}^{(s+1)}$, satisfying $\hat{\theta}^{(s+1)} = R(\hat{\theta}_L, \hat{K})$, where $\sum_i \Psi_L(Y_i^*, W_i;\hat{\theta}_L, \hat{K}) = 0$.

The general steps are as follows: starting with an initial guess $\hat{\theta}^{(0)}$, we iterate the following two steps for $s = 0, 1, 2, \ldots, S$ until $\hat{\theta}^{(s)}$ converges to the stationary distribution:

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

Reduction function. In practice, one of the challenges in implementing the PX-SEM algorithm is to find the reduction function associated with the L model. However, if we construct the L model such that the auxiliary parameter K does not affect the observed-data likelihood, that is $f_O(Y_i|X_i;\theta_L) = f_L(Y_i|X_i;\theta_L, K)$, then immediately, the reduction function becomes $R(\theta, K) = \theta$. As a result, PX-SEM can be simplified as follows:

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

[7] Note that the reduction function satisfies $R(\theta, K_0) = \theta$.

2. PX-M step: Update parameters by solving $\sum_i \Psi_L(Y_i^*, W_i; \hat{\theta}^{(s+1)}, \hat{K}) = 0$

Comparing the PX-M and M steps, we find that the M-step estimator is a constrained version of the PX-M-step estimator with the constraint $K = K_0$. Intuitively, when the E-step draws Y_i^* are generated under a guess $\hat{\theta}^{(s)}$ close enough to the true value, the M-step estimator is under the correct restriction, leading both the M-step and PX-M-step estimators to be consistent at that iteration, as indicated by (2) and (3).

However, when the guess $\hat{\theta}^{(s)}$ deviates significantly from the true value, causing the draws Y_i^* to violate certain model assumptions, we would expect the PX-SEM estimator to exhibit greater "robustness" due to extra flexibility brought by auxiliary parameter K. As shown in the following section, the likelihood-based PX-M step can achieve a larger pseudo-complete data likelihood improvement, which could further lead to a greater observed-data likelihood improvement compared to the M step.

3.2. Statistical Properties

This section focuses on the statistical properties of likelihood-based algorithms. We will first show that the parameter-expanded EM algorithm exhibits nonnegative improvement in the observed-data log-likelihood at each iteration. Next, for the stochastic version, PX-SEM, we will establish its asymptotic equivalence to an alternative SEM algorithm with a smaller expected fraction of missing information compared to the standard O model based SEM, which implies a faster global convergence rate and a smaller variance for the limiting stationary distribution in a semipositive definite order.

Convergence. Following Liu, Rubin, and Wu (1998), we now prove that PX-EM algorithm increases the observed-data likelihood in each iteration. The change in the observed-data log-likelihood between iterations $\sum_i \log f_O(Y_i|X_i; \hat{\theta}^{(s+1)}) - \sum_i \log f_O(Y_i|X_i; \hat{\theta}^{(s)})$ equals

$$\begin{aligned}&\sum_i \log f_L(Y_i|X_i; \hat{\theta}_L, \hat{K}) - \sum_i \log f_L(Y_i|X_i; \hat{\theta}^{(s)}, K_0)\\ &\geq Q(\hat{\theta}_L, \hat{K}|\hat{\theta}^{(s)}, K_0) - Q(\hat{\theta}^{(s)}, K_0|\hat{\theta}^{(s)}, K_0) \geq 0,\end{aligned}$$

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

The equality holds because of both condition (a): When $K = K_0$, two models coincide, meaning $f_O(Y_i|X_i; \hat{\theta}^{(s)}) = f_L(Y_i|X_i; \hat{\theta}^{(s)}, K_0)$, and condition (b): The reduction function exists, and thus by construction $f_O(Y_i|X_i; \hat{\theta}^{(s+1)}) = f_L(Y_i|X_i; \hat{\theta}_L, \hat{K})$. We then apply Gibbs' inequality. Finally, the definition of $\hat{\theta}_L$, which is $\hat{\theta}_L, \hat{K} \equiv \arg\max_{\theta,K} Q(\theta, K|\hat{\theta}^{(s)}, K_0)$, leads to a nonnegative change in observed-data likelihood. Notably, the result also implies that $(\hat{\theta}, K_0)$ is a fixed point of PX-EM, where $\hat{\theta}$ represents the MLE.[8]

As a final remark, the L model nesting the O model implies the following inequality:

$$\begin{aligned}&Q(\hat{\theta}_L, \hat{K}|\hat{\theta}^{(s)}, K_0) - Q(\hat{\theta}^{(s)}, K_0|\hat{\theta}^{(s)}, K_0)\\ &\geq Q(\hat{\theta}_{EM}^{(s+1)}, K_0|\hat{\theta}^{(s)}, K_0) - Q(\hat{\theta}^{(s)}, K_0|\hat{\theta}^{(s)}, K_0),\end{aligned}$$

where $\hat{\theta}_{EM}^{(s)} = \arg\max_\theta \sum_i \int \log f_O(Y_i, Y_i^*|X_i; \theta) f_L(Y_i^*|W_i; \hat{\theta}^{(s)}, K_0) dY_i^*$. Therefore, the parameter expansion technique can be intuitively interpreted as a way to improve the lower bound of the log-likelihood increment compared to the EM algorithm.

Asymptotic properties. We first characterize the dynamics of PX-SEM updates. Define Θ as the joint set of auxiliary and O model parameters, $\Theta \equiv [\theta; K]$. Accordingly, $\overline{\Theta} \equiv [\overline{\theta}; K_0]$ represents the vector of true values, and $\hat{\Theta} \equiv [\hat{\theta}; K_0]$ represents the MLE of the O model. Given any estimate $\hat{\Theta}^{(s)} = [\hat{\theta}^{(s)}; K_0]$ in the iteration s, PX-SEM generates the next update from a Markov process: $\sum_i \Psi_L(Y_i^*, W_i; \hat{\Theta}^{(s+1)}) = 0$, where $Y_i^* \sim f_L(Y_i^*|W_i; \hat{\Theta}^{(s)})$.[9,10] Expanding around $\hat{\Theta}$ and considering $\hat{\theta} \xrightarrow{p} \overline{\theta}$, as shown in detail in Appendix B, we have

$$\begin{aligned}(\hat{\Theta}^{(s+1)} - \hat{\Theta}) = (I - A^{-1}V)(\hat{\Theta}^{(s)} - \hat{\Theta}) + A^{-1}\epsilon^{(s)}\\ + o_p(N^{-(1/2)}),\end{aligned} \tag{4}$$

where, under the correct specification, $A = E\big(\Psi_L(Y_i^*, W_i; \overline{\Theta}) \Psi_L(Y_i^*, W_i; \overline{\Theta})'\big)$ is the L model-based complete-data information matrix, $V = E\big(\Psi_L(W_i; \overline{\Theta}) \Psi_L(W_i; \overline{\Theta})'\big)$ is the L model-based observed-data information matrix, $I - A^{-1}V$ is the expected fraction of missing information, and $\sqrt{N}\epsilon^{(s)} \xrightarrow{d} \mathcal{N}(\vec{0}, A - V)$.

The SEM iterations can be characterized in the same way:

$$\begin{aligned}(\hat{\theta}_{SEM}^{(s+1)} - \hat{\theta}) = \big(I - A_{\theta\theta}^{-1} V_{\theta\theta}\big)(\hat{\theta}_{SEM}^{(s)} - \hat{\theta})\\ + A_{\theta\theta}^{-1}\epsilon_\theta^{(s)} + o_p(N^{-(1/2)}),\end{aligned} \tag{5}$$

where $A_{\theta\theta}$ represents the O model-based complete-data information matrix, $V_{\theta\theta}$ represents the O model-based observed-data information matrix, $F_{SEM} \equiv I - A_{\theta\theta}^{-1} V_{\theta\theta}$ is the expected fraction of missing information, and $\sqrt{N}\epsilon_\theta^{(s)} \xrightarrow{d} \mathcal{N}(\vec{0}, A_{\theta\theta} - V_{\theta\theta})$. Moreover, PX-SEM and SEM dynamics are closely connected:

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

where $A_{\theta K} \equiv -\frac{\partial}{\partial K'}\big|_{\overline{\Theta}} E(\widetilde{\Psi}_L^\theta(Y_i^*, W_i; \Theta))$ and $A_{KK} \equiv -\frac{\partial}{\partial K'}\big|_{\overline{\Theta}} E(\widetilde{\Psi}_L^K(Y_i^*, W_i; \Theta))$.

We now present the main results of the asymptotic properties, building on Liu, Rubin, and Wu (1998) and Nielsen (2000), with detailed discussions provided in Appendix B.

Theorem 1. The PX-SEM iteration of $\hat{\theta}^{(s)}$ is asymptotically equivalent to SEM iteration with observed-data information matrix $V_{\theta\theta}$ and complete-data information matrix $A_{\theta\theta} - A_{\theta K} A_{KK}^{-1} A_{K\theta}$.

[8] In the moment-based PX-EM case, if the fixed point with $K = K_0$ exists, then it will satisfy $\sum_i \left(\int \Psi_O(Y_i^*, W_i; \hat{\theta}) f_O(Y_i^*|W_i; \hat{\theta}) dY_i^* \right) = 0$.

[9] Note that the E-step draws of PX-SEM are based on the O model under the guess $\hat{\theta}^{(s)}$. This is equivalent to making draws from the L model under the guess $\hat{\Theta}^{(s)} = [\hat{\theta}^{(s)}; K_0]$, due to condition (a).

[10] Appendix B shows that for any L model, an alternative L model can be found by reparameterization, which yields identical updates of θ with the reduction function $R(\theta, K) = \theta$.

Proof. Let H denote the inverse of matrix A, that is

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

where, by design, $H_{\theta\theta}^{-1} = A_{\theta\theta} - A_{\theta K}A_{KK}^{-1}A_{K\theta}$. Then the coefficient matrix $I - A^{-1}V$ and the asymptotic variance of the innovation term $A^{-1}\epsilon^{(s)}$ in (4) become:

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

$$A^{-1}(A - V)A^{-1}$$

$$= \begin{bmatrix} H_{\theta\theta}(H_{\theta\theta}^{-1} - V_{\theta\theta})H_{\theta\theta} & H_{\theta K} - H_{\theta\theta}V_{\theta\theta}H_{\theta K} \\ H_{K\theta} - H_{K\theta}V_{\theta\theta}H_{\theta\theta} & H_{KK} - H_{K\theta}V_{\theta\theta}H_{\theta K} \end{bmatrix}.$$

It becomes evident that the PX-SEM process of $\hat{\theta}^{(s+1)}$ is asymptotically equivalent to an alternative SEM dynamics, described by (6), which shares the same observed-data information matrix $V_{\theta\theta}$ as the standard SEM in (5), but replaces the original complete-data information matrix $A_{\theta\theta}$ by $H_{\theta\theta}^{-1} = A_{\theta\theta} - A_{\theta K}A_{KK}^{-1}A_{K\theta}$.

$$(\hat{\theta}^{(s+1)} - \hat{\theta}) = \big(I - H_{\theta\theta}V_{\theta\theta}\big)(\hat{\theta}^{(s)} - \hat{\theta}) + H_{\theta\theta}\widetilde{\epsilon}_{\theta}^{(s)} + o_p(N^{-(1/2)}), \tag{6}$$

where $\sqrt{N}\widetilde{\epsilon}_{\theta}^{(s)} \xrightarrow{d} \mathcal{N}(\vec{0}, H_{\theta\theta}^{-1} - V_{\theta\theta})$, and $F_{PX} \equiv I - H_{\theta\theta}V_{\theta\theta}$ is the expected fraction of missing information.[11] □

Since $H_{\theta\theta} = A_{\theta\theta}^{-1} + H_{\theta K}H_{KK}^{-1}H_{K\theta}$, under the condition that A is positive definite, $H_{\theta\theta} \geq A_{\theta\theta}^{-1}$ in a semipositive definite order, implying the largest eigenvalue of F_{PX} is no greater than the largest eigenvalue of F_{SEM}. Applying Theorem 1, this comparison in the expected fraction of missing information matrix between PX-SEM and SEM immediately implies the dominance of PX-SEM in convergence rate, as stated in Corollary 1.

Corollary 1. PX-SEM dominates SEM in global rate of convergence.

Proof. In Appendix B.2. □

Moreover, Nielsen (2000) provides conditions under which the SEM update is ergodic and characterizes the limiting stationary distribution, based on which Corollary 2 describes the limiting stationary distribution of PX-SEM and compares it with SEM.[12]

Corollary 2. The limiting stationary distribution of PX-SEM updates $\hat{\theta}^{(s)}$, conditional on W_i, is $\sqrt{N}(\hat{\theta}^{(s)} - \hat{\theta}) \xrightarrow{d} \mathcal{N}\Big(0, V_{\theta\theta}^{-1}\big(I-(I+F'_{PX})^{-1}\big)\Big)$, and unconditionally, is $\sqrt{N}(\hat{\theta}^{(s)} - \bar{\theta}) \xrightarrow{d} \mathcal{N}\Big(0, V_{\theta\theta}^{-1}\big(2I - (I + F'_{PX})^{-1}\big)\Big)$, with its variances being less than or equal to those of the standard O model-based SEM, that is, $V_{\theta\theta}^{-1}\big(2I - (I + F'_{PX})^{-1}\big) - V_{\theta\theta}^{-1}\big(2I - (I + F'_{SEM})^{-1}\big) = V_{\theta\theta}^{-1}\big(I - (I + F'_{PX})^{-1}\big) - V_{\theta\theta}^{-1}\big(I - (I + F'_{SEM})^{-1}\big) \leq 0$ in semipositive definite order.

[11] That $A - V$ being positive definite implies $H_{\theta\theta}^{-1} - V_{\theta\theta}$ being positive definite.

[12] The author thanks an anonymous referee for his/her encouragement to develop this result.

Proof. In Appendix B.2. □

Corollary 2 implies that PX-SEM updates exhibit smaller fluctuation along iterations in large samples. Moreover, since the final estimator is the average of the last S^0 iterations after convergence, $\hat{\theta} = \frac{1}{S^0}\sum_{S-S^0+1}^{S}\hat{\theta}^{(s)}$, which converges to the MLE as the number of iterations increases, PX-SEM and SEM estimators share the same asymptotic variance.[13]

When the M-step is moment-based, in general, convergence is not guaranteed. Under convergence, the speed does not necessarily dominate SEM. Indeed, Appendix A shows an example where moment-based PX-SEM underperforms SEM for some initial guesses.

However, moment-based PX-SEM may be the preferred choice in practice for at least two crucial reasons. First, in some cases, obtaining GMM estimators is much easier, such as in the quantile model discussed in Section 6. Since our final target is to reduce the computing time, we should consider not only the number of iterations but also the time spent in each iteration. Second, even if obtaining the MLE of the O model is feasible, restricting ourselves to a tractable MLE in the PX-M step can limit the flexibility in building the L model, negatively impacting the convergence rate. Appendix A shows an example of the toy model where the moment-based PX-SEM with a more flexible L model outperforms the likelihood-based PX-SEM, which uses a less flexible L model.

3.3. Implementation based on Linear Expansions

So far, we have shown that a new estimation method, PX-SEM, which combines the parameter expansion technique with the SEM algorithm, has attractive theoretical properties relative to ordinary SEM and the potential to achieve large computational gains.

However, the parameter expansion technique itself does not speak of the selection of the L model. On the one hand, all else being equal, a more flexible L model should improve the convergence rate. On the other hand, we also need to consider the time spent in each iteration to estimate the L model and convert it to the O model since our ultimate goal is to reduce the total computing time. Therefore, another contribution of this article is to propose a specific class of *linear* expansions, targeting the potential violation of zero-correlation assumptions, which can be generally applied to a wide range of models.

Considering an O model of the form: $Y_i = G(Y_i^*;\theta)$, $Y_i^* \sim F_O(\theta)$, where both $G(\cdot)$ and $F_O(\cdot)$ are known parametric functions up to unknown parameter θ, we propose the following linear expansion to Y_i^*:[14]

$$Y_i = G(Y_i^*;\theta), \quad Y_i^* = AY_i^\dagger, \quad Y_i^\dagger \sim F_O(\theta), \qquad (L\ Model)$$

$$\text{s.t. } G(Y_i^*;\theta) \overset{d}{=} G(Y_i^\dagger;\theta)\ (\textit{equally distributed}),$$

where the auxiliary parameter is given by $K = \text{vec}(A)$.[15]

[13] With a fixed S^0, the PX-SEM and SEM estimators will in general give rise to different asymptotic variances. Expressions for these variances are provided in Appendix B.

[14] In this expression, Y_i^* also includes error terms in the measure equation $G(\cdot)$.

[15] Extensions include unit-specific matrix A (Section 4) and adding exogenous regressor X_i (Section 5).

The expansion is straightforward. We assume that the latent variable $Y_i^\dagger$ follows the same distribution as the O model counterpart. However, the E-step draws Y_i^*, which directly contribute to the measurement equation and observable Y_i, result from an affine transformation applied to $Y_i^\dagger$. It is easy to check that when $A = I$, the L model coincides with the O model, whereas when $A \neq I$, it allows us to introduce linear correlations among elements of Y_i^*. The constraint ensures that the auxiliary parameters do not affect the observed-data likelihood, simplifying the reduction function to $R(\theta, K) = \theta$.[16]

To implement the PX-SEM algorithm, in the E-step, we draw Y_i^* from the O model based posterior distribution as discussed before. In the PX-M step, we leverage moment constraints or the distribution of $F_O(\theta)$ to pin down A and θ.

This method has the advantage that, despite the model of interest being nonlinear, the expansion is *linear in latent variables*, which are drawn from the E-step and treated as observables in the PX-M step. Thus, we can identify the auxiliary parameters separately through a relatively simple linear model, regardless of the specific form of $G(\cdot)$.

In the following sections, we discuss three applications: (a) dynamic factor models, (b) discrete choice models, and (c) quantile models, for which we propose PX-SEM algorithms based on linearly expanded models.

4. Dynamic Factor Models

The first type of model we discuss is the dynamic factor model (Geweke 1977). The appeal of this class of models is their ability to explain variation across multiple dimensions using fewer latent common factors. Applications span multiple fields, including topics in macroeconomics and finance, among others (Bai and Ng 2008; Stock and Watson 2006, 2011). While we will focus on a specific single-factor O model, it is worth noting that the same approach for implementing the PX-SEM algorithm can be applied to models with multiple latent factors. The O model to be estimated is as follows:

$$y_{it} = \lambda_i \nu_t + \epsilon_{it} \text{ and } \nu_t = \nu_{t-1} + u_t, \qquad (O\ Model)$$

where $\epsilon_{it} \overset{\text{iid}}{\sim} N(0, \sigma_i^2)$, $u_t \overset{\text{iid}}{\sim} N(0,1)$, $\nu_0 = 0$, and u_t is independent of ϵ_{it}.

The model contains a latent common factor ν_t that follows a Gaussian random walk. We observe N different measures, y_i, where $i = 1, \ldots, N$, over a total of T periods, with each measure associated with a distinct factor loading λ_i. The set of unknown parameters is denoted as $\theta \equiv (\lambda_1, \ldots, \lambda_N, \sigma_1, \ldots, \sigma_N)$.[17]

SEM. We first explain the SEM procedure. Let $\nu \equiv [\nu_1 \ \nu_2 \ \ldots \ \nu_T]'$. Starting with an initial guess $\hat{\theta}^{(0)}$, we iterate through the E-step and the M-step for $s = 0, 1, 2, \ldots, S$ until the convergence of $\hat{\theta}^{(s)}$ to the stationary distribution:

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

$$\hat{\lambda}_i = \Big(\sum_t \nu_t^2\Big)^{-1}\Big(\sum_t \nu_t y_{it}\Big) \text{ and } \hat{\sigma}_i = \widehat{\text{std}}(y_{it} - \hat{\lambda}_i \nu_i), \forall i$$

PX-SEM. To implement PX-SEM, we construct a simple L model as follows:

$$y_{it} = \lambda_i \nu_t + \epsilon_{it} \text{ and } \nu_t = \nu_{t-1} + u_t, \qquad (L\ Model)$$

where $\epsilon_{it} \overset{\text{iid}}{\sim} N(0, \sigma_i^2)$, $u_t \overset{\text{iid}}{\sim} N(0, k^2)$, $\nu_0 = 0$, and u_t is independent of ϵ_{it}.

This L model expands the O model by introducing an auxiliary parameter, k, allowing the variance of the persistent shock u_t to deviate from 1. Since k can always take the value of 1, making the two models coincide, the L model satisfies condition (a). Moreover, it is easy to verify that reduction function $R(\lambda_1, \ldots, \lambda_N, \sigma_1, \ldots, \sigma_N, k) = (\lambda_1 k, \ldots, \lambda_N k, \sigma_1, \ldots, \sigma_N)$ satisfies condition (b), that is $f_O(y_i; R(\theta, k)) = f_L(y_i; \theta, k)$.

With the L model specified and an initial guess $\hat{\theta}^{(0)}$, we iterate through the E-step and the PX-M step for $s = 0, 1, \ldots, S$ until the convergence of $\hat{\theta}^{(s)}$ to the stationary distribution:

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

$$\hat{\lambda}_{Li} = \Big(\sum_t \nu_t^2\Big)^{-1}\Big(\sum_t \nu_t y_{it}\Big) \text{ and}$$

$$\hat{\sigma}_{Li} = \widehat{\text{std}}(y_{it} - \hat{\lambda}_i \nu_i), \forall i; \text{ and } \hat{k} = \widehat{\text{std}}(\nu_t - \nu_{t-1})$$

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

The PX-M step estimation of the auxiliary parameter k is straightforward due to the separability of the log-likelihood function. Compared to SEM, the PX-SEM update $\hat{\theta}^{(s+1)}$ takes into account potential deviations from the assumption $k = 1$ in the O model. When the guess $\hat{\theta}^{(s)}$ is sufficiently close to the true value, we expect $\hat{k}$ to be close to 1, resulting in similar SEM and PX-SEM updates $\hat{\theta}^{(s+1)}$. However, when the guess $\hat{\theta}^{(s)}$ deviates significantly from the true value, leading to a violation of the assumption $k = 1$ in the E-step draws, PX-SEM adjusts the estimate accordingly. For instance, if $\hat{k}$ is greater than 1, it suggests scaling down the latent draws ν by a factor of k to ensure $\text{var}(\Delta\nu) = 1$ and scaling up λ_i by the same factor k to maintain the same log-likelihood for observed data.

Remark. It is easier to see the connection to the proposed linear expansion method after reparameterizing the L model. As detailed in Appendix C, we obtain an alternative expanded model with the reduction function $R(\theta, K) = \theta$, that is, $y_{it} = \lambda_i \nu_t + \epsilon_{it}$, $[\nu_1 \ \ldots \ \nu_T \ \epsilon_{i1} \ \ldots \ \epsilon_{iT}]' = A_i[\nu_1^* \ \ldots \ \nu_T^* \ \epsilon_{i1}^* \ \ldots \ \epsilon_{iT}^*]'$, $\nu_t^* = \nu_{t-1}^* + u_t$, where $A_i = \begin{bmatrix} kI_{T\times T} & 0_{T\times T} \\ \lambda_i(1-k)I_{T\times T} & I_{T\times T} \end{bmatrix}$, ϵ^*, u, and ν^* follow identical distributions to the O model counterparts. Thus, it is evident that the proposed L model belongs to the linear expansions with a specific constraint on matrix A_i: only contemporaneous correlations between (ν_t, ϵ_{it}) and $(\nu_t^*, \epsilon_{it}^*)$ are

[16] The constraint might not be necessary in applications where reduction functions are easy to find.

[17] The method can be easily adapted to models with (a) unknown persistence in the ν_t process, (b) multiple latent factors, (c) ϵ_{it} following an MA process, etc.

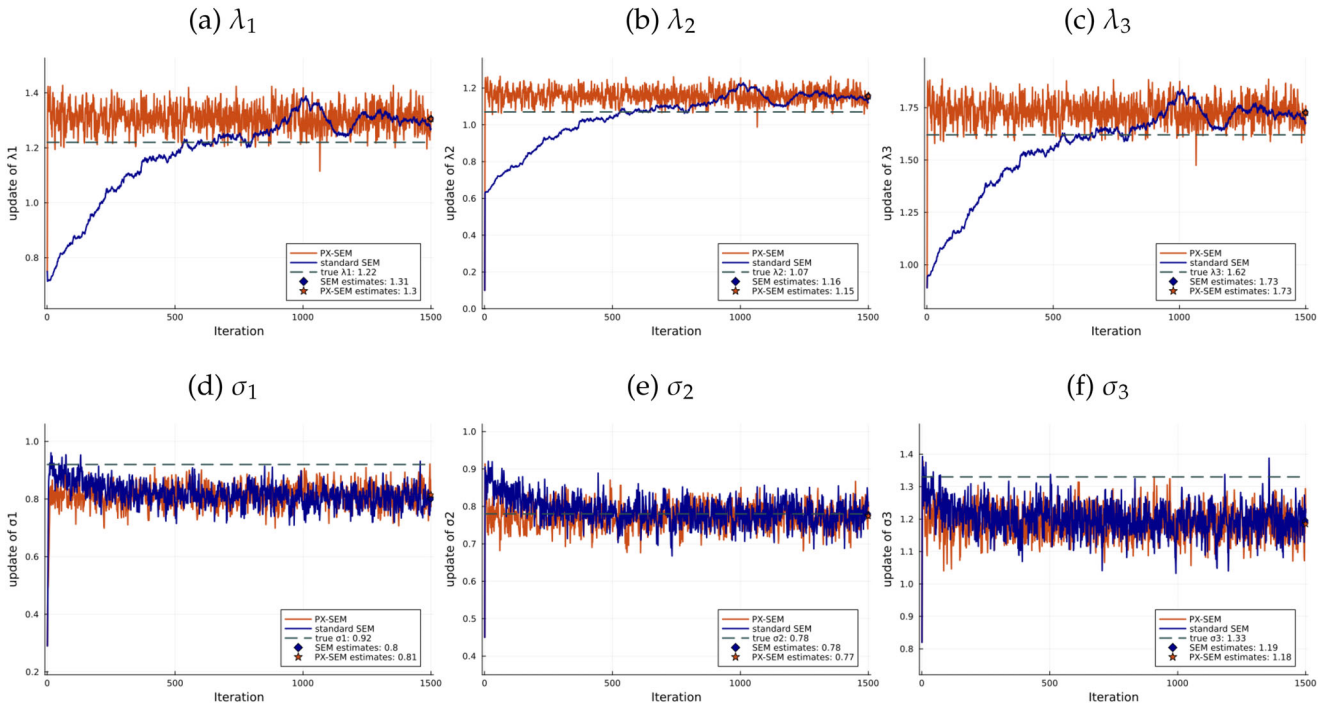

Figure 2. SEM and PX-SEM iterations of $\hat{\theta}^{(s)}$ from a random initial guess.
NOTE: Iterations of SEM (blue line) and PX-SEM (orange line) based on direct sampling, compared with the true value (green dashed line). SEM estimates (blue diamond) and PX-SEM estimates (orange star) are calculated as the average of the last 500 iterations. Random initial guess generated from a lognormal distribution. $N = 3, T = 200$.

allowed. Despite its advantage of the easy adaptation for various models and a negligible increase in computational burden due to the likelihood separability, in the other two applications, we will explore more flexible L models by relaxing constraints in the matrix A, such as allowing for correlations across periods, to achieve faster convergence.

Simulation Results. Figure 2 presents simulation results for a DGP where $\lambda = (1.22, 1.07, 1.62)$ and $\sigma = (0.92, 0.78, 1.33)$ with $N = 3$ and $T = 200$. The x-axis represents the number of iterations $s = 1, \ldots, 1500$, and the y-axis represents the M-step update $\hat{\theta}^{(s)}$. The blue line depicts the SEM trajectory, whereas the orange line depicts the PX-SEM trajectory. The horizontal green dashed line represents the true value. Starting from a randomly chosen initial guess $\hat{\theta}^{(0)}$, both SEM and PX-SEM updates move toward the true value and stabilize after several iterations. We use the average of the last 500 updates as the final estimate.

As shown in Figure 2, for all the parameters, PX-SEM converges almost immediately. However, for SEM, although it also converges relatively fast for $\sigma's$, a notable difference is observed in the case of $\lambda's$: it does not converge until 500 iterations.

Regarding volatilities of updates across iterations, Appendix D presents figures with longer trajectories, where we can observe that PX-SEM exhibits smaller volatilities. Appendix D also includes results for larger sample sizes and additional figures plotting cumulative computing time, revealing significant gains, especially for larger samples.

5. Discrete Choice Models

The second type of model we discuss is the random effects discrete choice model with persistent and transitory components. Discrete choice models are widely used in empirical research on various topics, including labor supply (Hyslop 1999) and consumer demand (Keane et al. 2013), among others. Distinguishing heterogeneity from persistence is of interest for many reasons, but the nonlinearity and the presence of latent variables complicate the estimation process.[18]

In this section, we develop PX-SEM algorithms for a group of discrete choice models with rich latent-variable structures, including time-invariant, persistent, and transitory components. Specifically, the O model is as follows:

$$\begin{aligned} y_{it} &= \mathbb{1}(z_{it} > 0), \qquad (O\ Model)\\ z_{it} &= \beta' x_{it} + \mu_i + \nu_{it} + \epsilon_{it},\\ \nu_{it} &= \rho \nu_{i,t-1} + u_{it}, \end{aligned}$$

where $\mu_i|x \overset{\text{iid}}{\sim} N(0, \sigma_\mu^2)$, $\nu_{i1}|x \overset{\text{iid}}{\sim} N(0,1)$, $u_{it}|x \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$, $\epsilon_{it}|x \overset{\text{iid}}{\sim} N(0,1)$; and $\mu_i, u_{it}, \epsilon_{it}$ are mutually independent.[19]

For each individual $i = 1, \ldots, N$ at period $t = 1, \ldots, T$, we observe a vector of independent variable x_{it} of dimension J and a binary (0-1) discrete dependent variable y_{it}, whereas z_{it}, individual effect μ_i, persistent component ν_{it}, and transitory component ϵ_{it} are latent. We denote the set of unknown parameters as θ, where $\theta \equiv (\beta, \sigma_\mu, \rho, \sigma_u)$.

SEM. Let $z_i \equiv [z_{i1} \ldots z_{iT}]'$, $\nu_i \equiv [\nu_{i1} \ldots \nu_{iT}]'$. From an initial guess $\hat{\theta}^{(0)}$, we iterate through E-step and M-step for $s = 0, 1, ..., S$ until $\hat{\theta}^{(s)}$ converges to the stationary distribution:

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

[18] Chen (2016) proposes a fixed effects EM estimator for a class of nonlinear panel data models.

[19] Appendix G presents two extensions: (a) Allowing for the dependence of μ_i and ν_{i1} on x_{i1}, and (b) Logit (with strategies for the quantile model in the next section).

2. M step: Update $\hat{\theta}^{(s+1)} = (\hat{\beta}^{(s+1)}, \hat{\sigma}_\mu^{(s+1)}, \hat{\rho}^{(s+1)}, \hat{\sigma}_u^{(s+1)})$

$$\hat{\beta}^{(s+1)} = \Big(\sum_i \sum_t x_{it}x_{it}'\Big)^{-1}\Big(\sum_i \sum_t x_{it}(z_{it} - \mu_i - \nu_{it})\Big),$$

$$\hat{\rho}^{(s+1)} = \Big(\sum_i \sum_t \nu_{i,t-1}\nu_{i,t-1}'\Big)^{-1}\Big(\sum_i \sum_t \nu_{i,t-1}\nu_{it}\Big),$$

$$\hat{\sigma}_\mu^{(s+1)} = \widehat{\text{std}}(\mu_i) \text{ and } \hat{\sigma}_u^{(s+1)} = \widehat{\text{std}}(\nu_{it} - \hat{\rho}^{(s+1)}\nu_{i,t-1}).$$

PX-SEM. One option for building the L model is to expand the O model to include only contemporaneous correlations, similar to the dynamic factor model in Section 4. Its advantage lies in the MLE being readily obtained in the PX-M step due to a separable likelihood. Appendix E provides the detailed steps and results of this approach. However, to achieve faster convergence, we now propose a more flexible L model.

Let us define $x_i \equiv [x_{i1}' \ \dots \ x_{iT}']'$, $\epsilon_i \equiv [\epsilon_{i1} \ \dots \ \epsilon_{iT}]'$, $\nu_i^* \equiv [\nu_{i1}^* \ \dots \ \nu_{iT}^*]'$, $\epsilon_i^* \equiv [\epsilon_{i1}^* \ \dots \ \epsilon_{iT}^*]'$, and $z_i^* \equiv [z_{i1}^* \ \dots \ z_{iT}^*]'$. We construct the following L model:

$$\begin{aligned} y_{it} &= \mathbb{1}(z_{it} > 0), \qquad (L\ Model)\\ z_{it} &= \gamma_t' x_i + \mu_i + \nu_{it} + \epsilon_{it},\\ [\mu_i \ \nu_i' \ \epsilon_i']' &= pA[\mu_i^* \ \nu_i^{*\prime} \ \epsilon_i^{*\prime}]' + Bx_i,\\ \nu_{it}^* &= \rho\nu_{i,t-1}^* + u_{it}, \end{aligned}$$

where $\mu_i^*|x \overset{\text{iid}}{\sim} N(0, \sigma_\mu^2)$, $\nu_{i1}^*|x \overset{\text{iid}}{\sim} N(0,1)$, $u_{it}|x \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$, $\epsilon_{it}^*|x \overset{\text{iid}}{\sim} N(0,1)$, and μ_i^*, u_{it}, ϵ_{it}^* are mutually independent; and subject to $\frac{1}{p} \times (CB + \gamma) = I_{T\times T} \otimes \beta'$, $CA\Sigma A'C' = C\Sigma C'$, and $p > 0$, where $C = [\overrightarrow{1}_{T\times 1} \ I_{T\times T} \ I_{T\times T}]$, $\Sigma = \text{var}([\mu_i^* \ \nu_i^{*\prime} \ \epsilon_i^{*\prime}]')$, $\gamma' \equiv [\gamma_1 \ \dots \ \gamma_T]$ and A is a lower triangular matrix with positive diagonal entries. Alongside θ from the O model, the L model contains a vector of auxiliary parameters $K \equiv [\text{vech}(A)', \text{vec}(B)', p]'$.

Following the linear expansion method, we introduce latent variables μ_i^*, ν_i^*, and ϵ_i^*, which follow the same distributions as their O model counterparts. However, the E-step draws for $[\mu_i \ \ \nu_i' \ \ \epsilon_i']$, given x_i, can result from an affine transformation applied to $[\mu_i^* \ \ \nu_i^{*\prime} \ \ \epsilon_i^{*\prime}]$, allowing for linear correlations among μ_i, ν_i, ϵ_i, and dependence on x_i. The scalar p permits scaling z_{it} and thus the deviation of $\text{var}(\epsilon_{it})$ from the value of 1. Hence, the L model satisfies condition (a): when $B = 0_{(2T+1)\times(J\times T)}$, $A = I_{(2T+1)\times(2T+1)}$, $p = 1$, the two models coincide $f_O(y_i, z_i, \mu_i, \nu_i|x_i;\theta) = f_L(y_i, z_i, \mu_i, \nu_i|x_i;\theta, A = I, B = \overrightarrow{0}, p = 1)$.

The L model has two key constraints: $\frac{1}{p} \times (CB+\gamma) = I_{T\times T} \otimes \beta'$ and $CA\Sigma A'C' = C\Sigma C'$. Beyond addressing identification, these constraints simplify the reduction function. Specifically, under these constraints, the L model can be written as $z_i = p\beta' x_{it} + pC[\mu_i^* \ \nu_i^{*\prime} \ \epsilon_i^{*\prime}]'$, implying no effect of auxiliary parameters p, A, and B on the conditional distribution of y_{it} given x_{it}. Therefore, regarding condition (b), we find a reduction function, $R(\theta, K) = \theta$, satisfying $f_O(y_i|x_i; R(\theta, K)) = f_L(y_i|x_i;\theta, K)$.

Finally, with the L model specified and an initial guess $\hat{\theta}^{(0)}$, we iterate through the following two steps for $s = 0, 1, \dots, S$ until $\hat{\theta}^{(s)}$ converges to the stationary distribution:

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

2. PX-M step:

   (a) L model estimation:

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

   (b) Reduction: $\hat{\theta}^{(s+1)} = R(\hat{\theta}_L, \hat{K}) = \hat{\theta}_L$

where $\Psi(\cdot)$ is a known function whose detailed specification is presented in Appendix F. Below, we list the moments involved in function $\Psi(\cdot)$:

$$\begin{aligned} p\beta: &\quad E(x_{it}(z_{it} - p\beta' x_{it})) = 0, \frac{1}{p} \times (CB + \gamma) = I_{T\times T} \otimes \beta'\\ B: &\quad E(x_i([\mu_i \ \nu_i' \ \epsilon_i'] - x_i'B')) = 0\\ \sigma_\mu, p, \Sigma, A: &\quad CA\Sigma A'C' = C\Sigma C', \text{and moment constraints on } \Sigma\\ \rho, \sigma_u: &\quad E(\nu_{i,t-1}^*(\nu_{it}^* - \rho\nu_{i,t-1}^*)) = 0, \text{var}(\nu_{it}^* - \rho\nu_{i,t-1}^*) = \sigma_u^2. \end{aligned}$$

Simulation Results. We conduct simulations to compare SEM and PX-SEM from a DGP with true parameter values: $\beta = [1.0; 0.5]$, $\sigma_\mu = 1.25$, $\rho = 0.7$, and $\sigma_u = 0.9$.

The initial guess is determined as follows: (a) $\hat{\beta}^{(0)}$ is the Probit regression coefficients of y_{it} on x_{it}, (b) Impose $\hat{\rho}^{(0)} = 1$, and the rest of the parameters are estimates of the linearly approximated model.[20] In the E-step, we employ a random-walk Metropolis-Hastings sampler with an acceptance rate controlled between 20% and 40%.

Figure 3 presents the estimation results of one simulation with $N = 5000$ and $T = 8$. Specifically, we plot the M-step updates $\hat{\theta}^{(s)}$ for 1000 iterations ($S = 1000$). The blue line depicts each update of SEM, while the orange one represents the PX-SEM updates. In this example, we might be interested in switching to SEM, which is likelihood-based, treating the PX-SEM estimate as an initial guess. Hence, we also present the results of a combined approach, where we run PX-SEM for 500 iterations and then continue with SEM for another 500 iterations, using the average of the last 250 PX-SEM iterations as the initial guess, as shown by the gray line.[21] The green dashed line indicates the true value. The final estimates are the average of the last 250 iterations ($S^0 = 250$), represented by the blue diamond for SEM, the orange star for PX-SEM, and the gray circle for PX-SEM+SEM.

From this comparison, it is clear that starting from the same initial guess, PX-SEM converges almost immediately (within 100 iterations). In contrast, SEM progresses much slower, especially for $\hat{\sigma}_\mu^{(s)}$, $\hat{\sigma}_u^{(s)}$, and $\hat{\rho}^{(s)}$, and does not converge within 1000 iterations. In terms of the combined approach, the variation across iterations significantly decreases after transitioning to SEM. However, since the final estimates are the average of the last 250 updates, the difference between PX-SEM and PX-SEM+SEM is small.[22]

[20] We approximate the model as follows $y_{it} = \Phi(\beta' x_{it} + \mu_i + \nu_{it}) + \eta_{it} \approx 0.5 + 0.25(\beta' x_{it} + \mu_i + \nu_{it}) + \eta_{it}$.

[21] We could also endogenize the switching procedure by using metrics like the distance between $\hat{K}$ and K_0 or the likelihood difference between the L model and the O model to guide our transition to SEM.

[22] Whether PX-SEM requires more iterations in this example due to its higher volatility, impacting total computing time, is beyond this article's scope, especially considering the PX-SEM+SEM option.

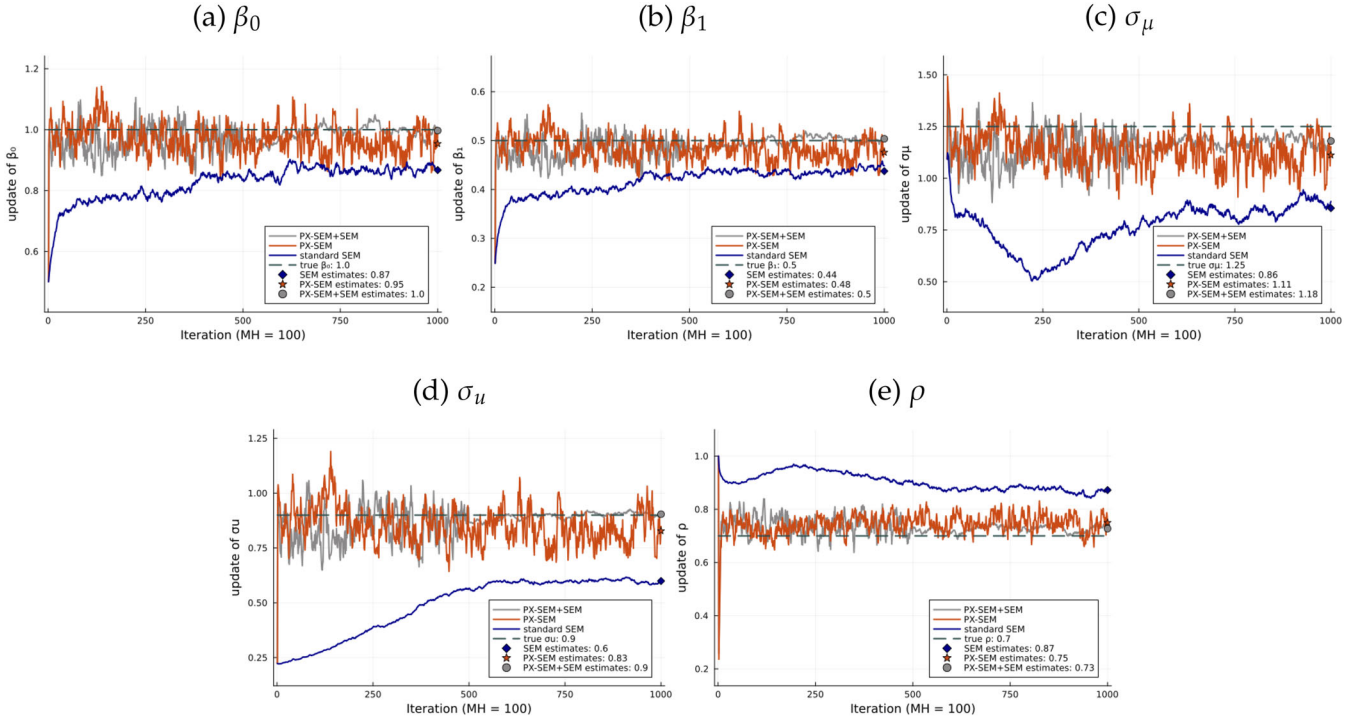

Figure 3. SEM and PX-SEM iterations of $\hat{\theta}^{(s)}$ from an informed guess.
NOTE: Iterations of SEM (blue line), PX-SEM (orange line), and PX-SEM+SEM (gray line) based on 100 MH draws, compared with the true value (green dashed line). Estimates of SEM (blue diamond), PX-SEM (orange star), and PX-SEM+SEM (gray circle) are the average of the last 250 iterations. Informed initial guess.

Appendix L provides additional figures where the x-axis is the cumulative computing time. The gain is significant: SEM takes approximately 3000 sec to run 1000 iterations without converging, while PX-SEM converges almost immediately.[23]

Appendix H compares algorithms based on random initial guesses. Researchers often run SEM algorithms from various initial guesses and choose one based on specific criteria (e.g., the likelihood value) to avoid obtaining a local maximum, given that getting a "good" initial guess can be challenging. Appendix H shows that the dominance of PX-SEM in convergence rate remains under random initial guesses. Given that this type of exercise is often performed repeatedly in practice, the time saved could be substantial.[24]

Finally, Appendix M provides the overall trajectories of SEM and PX-SEM over iterations. Specifically, we conduct 40 simulations using the same DGP, each estimated under a different set of initial guesses shared by both SEM and PX-SEM. For each parameter, we examine the distribution of updates across 40 trajectories at each specific iteration and how this distribution evolves over the iterations for SEM and PX-SEM, respectively. We reach the same conclusion: PX-SEM significantly improves algorithmic efficiency.

6. Quantile Models

The final type of model for which we consider a PX-SEM approach is the persistent-transitory dynamic quantile models with individual effects, as proposed by Arellano, Blundell, and Bonhomme (2017) (referred to as ABB hereafter). The ABB model does not impose functional-form restrictions on the distributions of individual effects, transitory shocks, or conditional distributions of the persistent component. Indeed, the flexible dynamics of the persistent component allow for attractive features such as nonlinear persistence, meaning that the persistence could vary with the size of shocks and accumulated history, which is shown to be empirically prominent in earning dynamics. The model has also been applied to other topics including firm and health dynamics.

Specifically, we focus on the ABB baseline model with an additive fixed effect, discussed in their Appendix.[25] Denote the τth conditional quantile of ν_{it} given $\nu_{i,t-1}$ as $Q_\nu(\nu_{i,t-1}, \tau)$ for each $\tau \in (0, 1)$. The O model to be estimated is as follows:

$$\begin{aligned} y_{it} &= \mu_i + \nu_{it} + \epsilon_{it}, \qquad (O\ Model) \\ \nu_{it} &= Q_\nu(\nu_{i,t-1}, u_{it}), \\ (u_{it}|\mu_i, u_{i,t-1}, u_{i,t-2}, \ldots) &\overset{\text{iid}}{\sim} \text{Uniform}(0,1),\ t = 2, \ldots, T, \end{aligned}$$

where ϵ_{it} has zero mean, iid over time, and independent of $\nu_i \equiv [\nu_{i1}\ \nu_{i2}\ \ldots\ \nu_{iT}]'$ and μ_i. Individual effect μ_i is assumed to be independent of $\epsilon_i \equiv [\epsilon_{i1}\ \epsilon_{i2}\ \ldots\ \epsilon_{iT}]'$ and ν_i.

To estimate this model, we follow Arellano, Blundell, and Bonhomme (2017) and empirically specify the quantile function of ν_{it} given $\nu_{i,t-1}$, $Q_\nu(\nu_{i,t-1}, \tau)$, the quantile function of ϵ_{it}, $Q_\epsilon(\tau)$, the quantile function of ν_{i1}, $Q_{\nu_1}(\tau)$, and the quantile

[23] The results are obtained using a Mac Mini (M1, 2020) with a single processor core. We apply the Metropolis-Hastings algorithm for the E-step, with the first 100 iterations designated as a burn-in phase.

[24] Appendix H also presents simulation results with more iterations and different sample sizes.

[25] In practice, standard SEM generally performs well in estimating the ABB baseline model without the fixed effect. But it is challenging when a fixed effect is included. We also remove age effects.

function of μ_i, $Q_\mu(\tau)$, as follows:

$$Q_v(v_{i,t-1},\tau)=\sum_{h=0}^{H}\gamma_h^Q(\tau)\varphi_h(v_{i,t-1}),$$
$$Q_\epsilon(\tau)=\gamma^\epsilon(\tau),\quad Q_{v_1}(\tau)=\gamma^{v_1}(\tau),\quad Q_\mu(\tau)=\gamma^\mu(\tau),$$

where $\varphi_h(\cdot)$ is Hermite polynomials of order h and $\gamma(\cdot)'s$ are functions to be estimated.

Arellano, Blundell, and Bonhomme (2017) exploit a variation of SEM for estimation, where the M-step involves a series of quantile regressions instead of likelihood optimization for computational convenience. We first explain their procedures. Let θ denote the set of unknown parameters, including $\gamma_k^Q(\tau)$, $\gamma^\epsilon(\tau)$, $\gamma^{v_1}(\tau)$, and $\gamma^\mu(\tau)$.[26] With an initial guess $\hat\theta^{(0)}$, we iterate through the following two steps until $\hat\theta^{(s)}$ converges to the stationary distribution:

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

$$\hat\gamma^Q(\tau)=\arg\min_{\gamma_0^Q,\ldots,\gamma_H^Q}\sum_{i=1}^{N}\sum_{t=2}^{T}\rho_\tau\Big(v_{it}-\sum_{h=0}^{H}\gamma_h^Q\varphi_h(v_{i,t-1})\Big),$$
$$\hat\gamma^\mu(\tau)=\arg\min_{\gamma^\mu}\sum_{i=1}^{N}\rho_\tau\big(\mu_i-\gamma^\mu\big),$$
$$\hat\gamma^\epsilon(\tau)=\arg\min_{\gamma^\epsilon}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_\tau\big(\epsilon_{it}-\gamma^\epsilon\big),$$
$$\hat\gamma^{v_1}(\tau)=\arg\min_{\gamma^{v_1}}\sum_{i=1}^{N}\rho_\tau\big(v_{i1}-\gamma^{v_1}\big),$$

where $\rho_\tau(u)=u(\tau-\mathbb{1}(u\le 0))$ is the check function.

PX-SEM. We expand the O model linearly targeting the correlations among μ_i, v_i, and ϵ_i. Define $v_i^*\equiv[v_{i1}^*\ \cdots\ v_{iT}^*]'$, $\epsilon_i^*\equiv[\epsilon_{i1}^*\ \cdots\ \epsilon_{iT}^*]'$. We build the following L model:

$$\begin{aligned}y_{it}&=\mu_i+v_{it}+\epsilon_{it},\qquad (L\ Model)\\ [\mu_i\ v_i'\ \epsilon_i']'&=A[\mu_i^*\ v_i^{*\prime}\ \epsilon_i^{*\prime}]',\\ v_{it}^*&=Q_v(v_{i,t-1}^*,u_{it}),\\ (u_{it}|\mu_i^*,u_{i,t-1},u_{i,t-2},\ldots)&\overset{\text{iid}}{\sim}\text{Uniform}(0,1),\ t=2,\ldots,T,\end{aligned}$$

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

Consistent with the linear expansion method, E-step draws $[\mu_i\ v_i'\ \epsilon_i']'$ are assumed to be outcomes of affine transformations through matrix A of $[\mu_i^*\ \ v_i^{*\prime}\ \ \epsilon_i^{*\prime}]'$, which follow identical distributions as their O model counterparts. When $A=I$, two models coincide, satisfying condition (a). Moreover, with the constraint $CA=C$, the L model becomes $y_i=C[\mu_i^*\ v_i^{*\prime}\ \epsilon_i^{*\prime}]'$, implying no effect of K on the observed-data likelihood. Thus, regarding condition (b), the reduction function is $R(\theta,K)=\theta$.

Finally, with this L model and an initial guess $\hat\theta^{(0)}$, we iterate through the following two steps for $s=0,1,,\ldots,S$ until $\hat\theta^{(s)}$ converges to the 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) L model estimation:
   $$\hat\theta_L,\hat K=\arg\min_{\theta,K}\sum_i\Psi(\theta,K;y_i,\mu_i,v_i)$$
   (b) Reduction: $\hat\theta^{(s+1)}=R(\hat\theta_L,\hat K)=\hat\theta_L$

where $\Psi(\cdot)$ is a known function to be discussed in the following paragraphs.

The inclusion of matrix A adds complexity to joint estimation due to infeasible separate quantile regressions as in the SEM M-step and the involvement of many more parameters.[27] Thus, we employ two strategies: adding extra constraints on the entries of matrix A and sequential estimation.

Regarding the extra constraints, we assume that ϵ_{it} is orthogonal to μ_i^*, ϵ_{i1}^*,..., $\epsilon_{i,t-1}^*$, for $t=2,\ldots,T$, and the coefficient of ϵ_{it}^* with respect to ϵ_{it} is homogenous across all periods. The advantage of doing so is that, by exploiting moment conditions including zero correlation among μ_i^*, v_i^*, and ϵ_i^*, and v_{it}^* following the first-order Markov process as well as ϵ_{it}^* being iid, we can separately estimate matrix A through constrained GMM while restricting the number of unknowns to only two. This further facilitates the sequential estimation strategy. Once obtaining $\hat A$, we estimate θ through the same series of quantile regressions as in SEM using $[\hat\mu_i^*\ \hat v_i^{*\prime}\ \hat\epsilon_i^{*\prime}]'=\hat A^{-1}[\mu_i\ v_i'\ \epsilon_i']'$. Appendix I provides a detailed discussion of the estimation process.[28]

Simulation Results. We simulate from the following DGP ($N=5000, T=6$):

$$\begin{aligned}y_{it}&=\mu_i+v_{it}+\epsilon_{it},\qquad (7)\\ v_{it}&=\rho_v v_{i,t-1}+(\sigma_{vt0}+\sigma_{vt1}v_{i,t-1}^2)\nu_{it},\end{aligned}$$

where $\mu_i\overset{\text{iid}}{\sim}N(0,\sigma_\mu^2)$, $v_{i1}\overset{\text{iid}}{\sim}N(0,\sigma_{v1}^2)$, $v_{it}\overset{\text{iid}}{\sim}N(0,1)$, $\epsilon_{it}\overset{\text{iid}}{\sim}N(0,\sigma_\epsilon^2)$, and $\mu_i, v_{i1}, \nu_{it}, \epsilon_{it}$ are mutually independent.

We present results for a persistent-transitory process without time-invariant heterogeneity and heteroscedasticity in the persistent shock by imposing $\mu_i=0$ and $\sigma_{vt1}=0$. The other parameter values are $\rho_v=0.8, \sigma_{vt0}=0.15, \sigma_{vt1}=0, \sigma_{v1}^2=0.15, \sigma_\epsilon^2=0.05$. Appendix J provides simulation results for a DGP with time-invariant heterogeneity and heteroscedasticity and another DGP, which is based on a flexible quantile model.

With the simulated data, we estimate the quantile model *with* time-invariant heterogeneity, as specified previously (the O model), assuming no knowledge of the distribution family. We set the initial guess by estimating the canonical random-walk permanent-transitory model, with details explained in

[26] 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 a parametric model. Please refer to Appendix B in Arellano, Blundell, and Bonhomme (2017) for more details.

[27] In the discrete choice model, the matrix A and other auxiliary parameters can be easily estimated in the PX-M step by focusing solely on the first two moments due to the normality assumption.

[28] Similar strategies are used to estimate a Logit model in Appendix G.

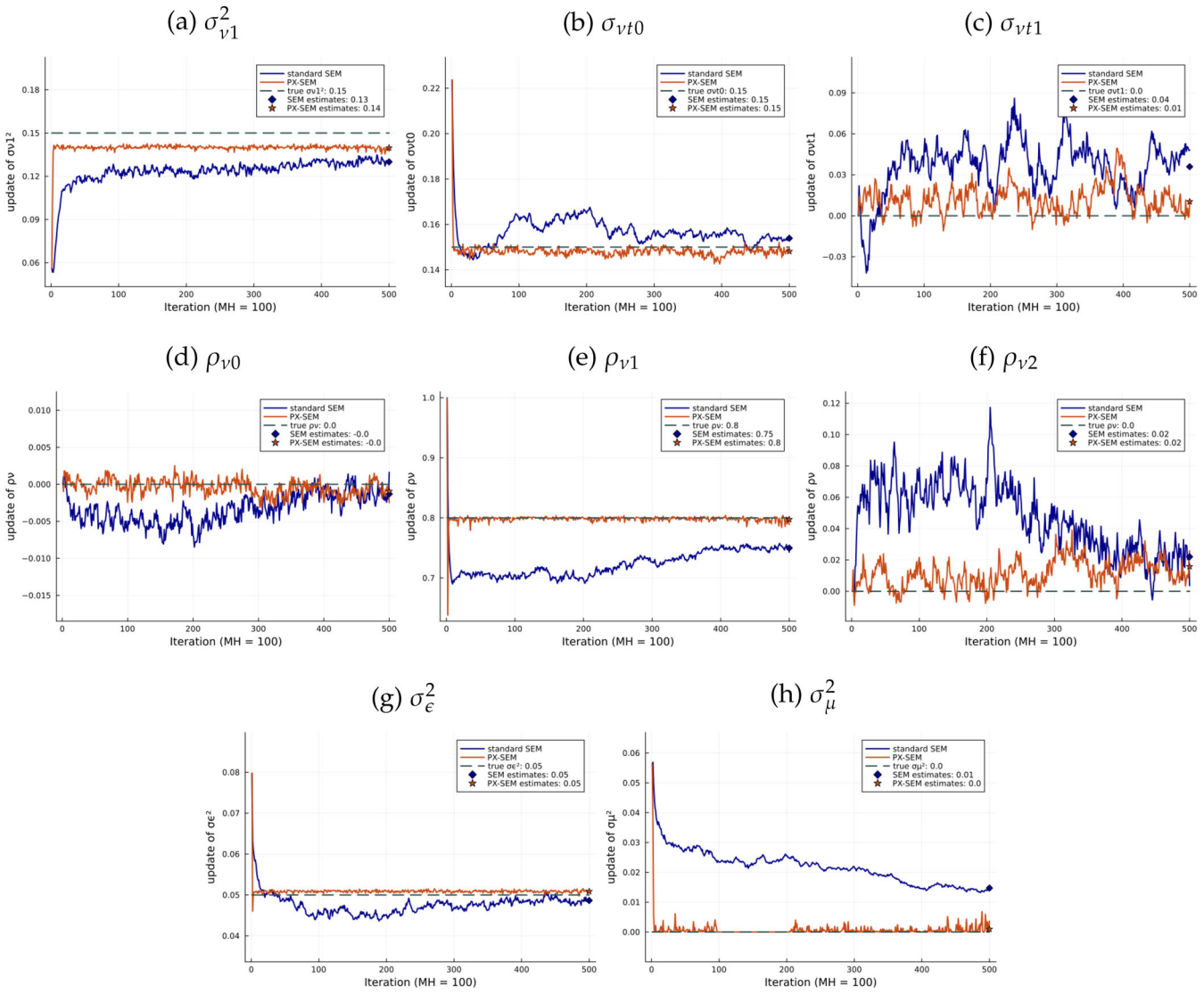

Figure 4. SEM and PX-SEM iterations, $\mu_i = 0$.
NOTE: Iterations of SEM (blue solid line) and PX-SEM (orange solid line) based on 100 MH draws, compared with the true value (green dashed line). In each iteration, we estimate the parametric model, (7), using E-step draws μ, ν, ϵ for SEM and "corrected" draws $\hat{\mu}^*$, $\hat{\nu}^*$, $\hat{\epsilon}^*$ for PX-SEM. These estimates are only used for visualizing the convergence and are not directly involved in any algorithm. SEM estimates (blue diamond) and PX-SEM estimates (orange star) are both calculated as the average of the last 100 iterations. Informed initial guess.

Appendix J. Finally, the highest order of Hermite polynomials for the empirical specification of the ν_{it} dynamics, H, is set to two.

Figure 4 presents the results. To provide clearer visualization, instead of plotting the updates of raw parameters in the quantile model directly (due to their large quantity), we plot the iterations of estimated parameter values in the parametric model, (7). Specifically, in each iteration, we estimate the parametric model using E-step draws (μ, ν, and ϵ) for SEM and "corrected" draws ($\hat{\mu}^*$, $\hat{\nu}^*$, and $\hat{\epsilon}^*$) for PX-SEM. Importantly, these estimates are only for visualizing convergence and are not directly involved in the algorithm updating procedure. Consistent with previous exercises, PX-SEM exhibits rapid convergence for all parameters, whereas SEM converges much more slowly.

Appendix L provides complementary figures to Figure 4, with cumulative computing time as the x-axis, showing significant time gain: SEM takes over 5500 sec for 500 iterations without clear convergence, whereas PX-SEM converges almost immediately.[29] Appendix M presents the overall trajectories of 40 iterations. Finally, Appendix K shows simulation results for different sample sizes.

After 500 iterations, averaging the last 100 updates as temporary estimates, we simulate from the estimated models and compare the model fit, focusing on the persistence of the ν_{it} dynamics, $\frac{\partial \nu_{it}}{\partial \nu_{i,t-1}}(\nu_{i,t-1}, \tau)$, one of the characteristics of interest (Arellano, Blundell, and Bonhomme 2017). Figure 5 displays heatmaps showing the absolute distance between the estimated persistence and true persistence at different levels of the shock τ and $\nu_{i,t-1}$ for SEM and PX-SEM, respectively. Overall, PX-SEM shows a better model fit.

[29] The results are obtained using a Mac Mini (M1, 2020) with a single processor core. We apply the Metropolis-Hastings algorithm for the E-step, with the first 100 iterations designated as a burn-in phase.

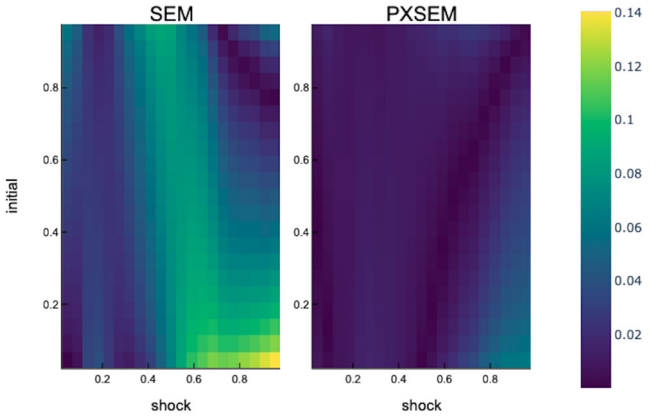

Figure 5. Distance between estimated persistence and true persistence, $\mu_i = 0$. NOTE: The absolute distance between the estimated ν_t persistence (based on the average of the last 100 iterations) and true persistence for each level of the shock τ and $\nu_{i,t-1}$.

7. Conclusions

This article introduces new estimation algorithms for dynamic panel data models with latent variables. By combining the parameter expansion ideas with the SEM algorithm, we develop the PX-SEM algorithm, which could facilitate convergence in models with a large space of latent variables by improving algorithmic efficiency.

Sharing the same E-step as SEM, PX-SEM differs in the M-step. Instead of estimating the original model (the O model), the M-step of PX-SEM requires estimating an expanded model (the L model). Effectively, we propose new estimators for the pseudo-data within iterations, accounting for the misspecification of the O model for draws based on parameter values far from the truth. Thus, PX-SEM can leverage additional model information to effectively "correct" the M-step updates in progressing to more accurate ones.

Moreover, the article proposes a method for constructing the L model through linear expansion and presents new PX-SEM-based estimation algorithms for three types of dynamic panel data models: factor models, discrete choice models, and quantile models.

Regarding statistical properties, we establish the asymptotic equivalence of the likelihood-based PX-SEM to an alternative SEM with a smaller expected fraction of missing information compared to the standard O model based SEM, implying a faster global convergence rate and a smaller variance for the limiting stationary distribution. Finally, simulations show that PX-SEM can significantly improve the algorithmic efficiency relative to SEM.

Supplementary Materials

The online supplement consists of the following appendices. Appendix A presents illustrative figures for the intuition behind PX-SEM and comparisons among different L models and M-step estimators using the toy model. Appendix B provides a detailed proof for Section 3. Appendix C explains the equivalence through reparameterization among L models. Appendices E, F, and G discuss alternative L models, detailed L model estimation procedures, and PX-SEM methods applied to two extensions for the discrete choice model in Section 5. Appendix I provides detailed L model estimation procedures for the quantile model in Section 6. Appendices D, H, J – M present more simulation results for the three types of models discussed in Sections 4-6, with more iterations, different sample sizes, cumulative computing time, different initial guesses, and overall trajectories.

Acknowledgments

This work is based on Chapter 2 of my PhD thesis at CEMFI, which received the Enrique Fuentes Quintana Funcas Award in Economics, Finance and Business 2021-2022. I am deeply grateful to Manuel Arellano for his invaluable support and advice. I also thank Martin Almuzara, Dante Amengual, Dmitry Arkhangelsky, Orazio Attanasio, Richard Blundell, Stéphane Bonhomme, Micole De Vera, Jose Gutierrez, Pedro Mira, Josep Pijoan-Mas, Enrique Sentana, Liyang Sun, and seminar participants at IE university, CEMFI, International Panel Data Conference, EEA-ESEM, and SAEe meetings for valuable comments and suggestions. Two anonymous referees and an Associate Editor have helped greatly improve the article. All errors are my sole responsibility.

Disclosure Statement

The author reports there are no competing interests to declare.

Funding

Grants PID2022-143184NA-I00 funded by MCIU/AEI/10.13039/501100011033 and by FEDER, UE; BES-2017-082506 funded by MCIU/AEI/10.13039/501100011033 and by "ESF Investing in your future"; and MDM-2016-0684 funded by MCIU/AEI/10.13039/501100011033 are greatly acknowledged.

ORCID

Siqi Wei http://orcid.org/0000-0002-8048-3609

References

Arcidiacono, P., and Jones, J. B. (2003), "Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm," *Econometrica*, 71, 933–946. [2]

Arellano, M., Blundell, R., and Bonhomme, S. (2017), "Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework," *Econometrica*, 85, 693–734. [10,11,12]

Arellano, M., Blundell, R., Bonhomme, S., and Light, J. (2023), "Heterogeneity of Consumption Responses to Income Shocks in the Presence of Nonlinear Persistence," *Journal of Econometrics*, 240, 105449. [1,2]

Arellano, M., and Bonhomme, S. (2016), "Nonlinear Panel Data Estimation via Quantile Regressions," *The Econometrics Journal*, 19, C61–C94. [2]

——— (2017), "Nonlinear Panel Data Methods for Dynamic Heterogeneous Agent Models," *Annual Review of Economics*, 9, 471–496. [1]

Bai, J., and Ng, S. (2008), *Large Dimensional Factor Analysis. Foundations and Trends® in Econometrics* (Vol. 3), pp. 89–163, Hanover, MA: Now Publishers. [7]

Chen, M. (2016), "Estimation of Nonlinear Panel Models with Multiple Unobserved Effects," Working Paper, Department of Economics, University of Warwick. [2,8]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1–22. [1,2]

Diebolt, J., and Celeux, G. (1993), "Asymptotic Properties of a Stochastic EM Algorithm for Estimating Mixing Proportions," *Stochastic Models*, 9, 599–613. [1,2]

Geweke, J. (1977), "The Dynamic Factor Analysis of Economic Time Series," in *Latent Variables in Socio-Economic Models*, eds. D. J. Aigner and A. S. Goldberger, Amsterdam: North-Holland. [7]

Hyslop, D. R. (1999), "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women," *Econometrica*, 67, 1255–1294. [8]

Keane, M. P. (2013), "Panel Data Discrete Choice Models of Consumer Demand," in *The Oxford Handbook of Panel Data*, ed. B. H. Baltagi, pp. 548–582, Oxford: Oxford University Press. [8]

Lavielle, M., and Meza, C. (2007), "A Parameter Expansion Version of the SAEM Algorithm," *Statistics and Computing*, 17, 121–130. [2]

Liu, C., Rubin, D. B., and Wu, Y. N. (1998), "Parameter Expansion to Accelerate EM: The px-em Algorithm," *Biometrika*, 85, 755–770. [1,2,5]

Liu, J. S., and Wu, Y. N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274. [2]

Nielsen, S. F. (2000), "The Stochastic EM Algorithm: Estimation and Asymptotic Results," *Bernoulli*, 6, 457–489. [2,5,6]

Pastorello, S., Patilea, V., and Renault, E. (2003), "Iterative and Recursive Estimation in Structural Nonadaptive Models," *Journal of Business & Economic Statistics*, 21, 449–509. [2]

Stock, J. H., and Watson, M. W. (2006), "Forecasting with Many Predictors," *Handbook of Economic Forecasting*, 1, 515–554. [7]

——— (2011), "Dynamic Factor Models," in *Oxford Handbook of Economic Forecasting*, eds. Michael P. Clements and David F. Hendry, Oxford: Oxford University Press. [7]

Wei, G. C., and Tanner, M. A. (1990), "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American statistical Association*, 85, 699–704. [2]

Wei, S. (2022), "Income, Employment and Health Risks of Older Workers," *Documentos de Trabajo (CEMFI)*, (5), 1. [1]

Wu, C. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103. [2]