Comments on "How Variational Are Your Earnings?"
by Neele Balke, Stéphane Bonhomme, and Thibaut Lamadon

Manuel Arellano
CEMFI

CostasFest: A Celebratory Conference in Honor of Costas Meghir

Yale University, September 12-13, 2025

Introduction

- In recent years, new facts on income dynamics emerged, both from large admin data sets and from surveys using new techniques that reveal different aspects of data.
- Skewness matters: predictive income distributions are negatively skewed for the rich, positively skewed for the poor, and become more negatively skewed in downturns.
- Persistence is nonlinear and varies greatly depending on the past income level and the magnitude of current shocks.
- These facts matter for understanding the effects of welfare and macro policies.
- As a result, models for macro and labor policy analysis increasingly incorporate nonlinear earnings processes, creating a parallel need for new econometric methods.

Introduction (continued)

- It is usual to begin with a persistent/transitory formulation for (residualized) earnings:

$$y_{it} = z_{it} + e_{it}$$

  where z_{it} and e_{it} are flexibly specified Markovian and independent processes.

- Then use an EM algorithm that alternates between MCMC sampling from the posterior $p\left(z_{1:T} | y_{1:T}\right)$ and updating parameter estimates via the method of moments.

- This procedure works well on short survey panels or moderately large admin datasets (e.g. $n = 200,000$ and $T = 6$).

- But its feasibility for larger cross-sections, and especially for longer panels, is unclear.

- The goal of BBL is to examine the performance of an alternative approach using variational inference; computationally and in its ability to recover true parameters.

- This project initiates a very useful conversation between different traditions.

The variational inference (VI) approach

- The EM algorithm maximizes the expected complete log-likelihood by iteratively computing a sample counterpart of the fixed point:

$$\theta_0 = \arg\max_{\theta} E\left(\int_z \left[\ln f_\theta\left(y_{1:T}, z_{1:T}\right)\right] p_{\theta_0}\left(z_{1:T} | y_{1:T}\right) dz\right).$$

- VI selects a family of distributions $q_\lambda\left(z_{1:T}; y_{1:T}\right)$ to approximate the true posterior, and then solves a sample counterpart of the following optimization problem:

$$\left(\theta^*, \lambda^*\right) = \arg\max_{\theta, \lambda} E\left(\int_z \left[\ln f_\theta\left(y_{1:T}, z_{1:T}\right)\right] q_\lambda\left(z_{1:T}; y_{1:T}\right) dz - \int_z \left[\ln fq_\lambda\left(z_{1:T}; y_{1:T}\right)\right] q_\lambda\left(z_{1:T}; y_{1:T}\right) dz\right).$$

- In this way, $q_{\lambda^*}\left(z_{1:T}; y_{1:T}\right)$ is the best approximation to $p_{\theta_0}\left(z_{1:T} | y_{1:T}\right)$ in a KL sense.

- The goal is to select $q_\lambda\left(z_{1:T}; y_{1:T}\right)$ (and $f_\theta\left(y_{1:T}, z_{1:T}\right)$) in such a way that optimization is simple and the approximation is accurate.

Summary

- BBL consider a model for $f_\theta(y_{1:T}, z_{1:T})$ of the form

$$z_t \mid z_{1:t-1} \sim \mathcal{N}\left[\mu(z_{t-1}), \sigma(z_{t-1})\right].$$

- The distributions of z_1 and e_t, and $\mu(z_{t-1})$ and $\sigma(z_{t-1})$ are flexibly specified.
- The approximate posteriors q_λ are multivariate normal with unrestricted, tridiagonal, or diagonal precision matrix, and a flexible family that exploits the Markov property.
- A reparameterization trick enables gradient-based optimization.
- In their simulations, Gaussian q_λs do well even when the true posterior is not Gaussian, as long as q_λ does not impose independence or ignores transitory shocks.
- A Jacobian-based transformation of the Gaussian VI posterior can allow for non-normal skewness and kurtosis, while preserving differentiability.
- BBL discuss "economically meaningful" diagnostics: implied risk premia and a discounted mixture density, combining densities for future z conditioned on initial z.
- Although the model is conditionally Gaussian, so that shocks one-period-ahead are symmetric, it can generate conditional skewness over multiple periods ahead.
- The discounted density estimated with PSID data exhibits ABB nonlinear persistence.

Consistency: prediction vs parameter estimation

- Often, variational methods are used in contexts of prediction or classification where structural interpretation of the latent variables is not central.

- This is not the case of persistent/transitory income models.

- Functionals of the latent persistent and transitory processes are of key interest, so we would like to have estimates with statistical guarantees.

- It is of interest to know what combinations of families of likelihood models $f_\theta\left(y_{1:T}, z_{1:T}\right)$ and variational posteriors $q_\lambda\left(z_{1:T}; y_{1:T}\right)$ are such that VI estimates are potentially consistent; that is, situations where

$$\theta^* = \theta_0$$

  for all or a subset of the likelihood parameters.

- Classic PML consistency results of conditional mean parameters, including the correlated random effects Mundlack model, and the existing results in the VI literature are an encouraging starting point (eg Westling & McCormick 2019).

Location-scale models

- Motivated by the challenge of estimating ABB-like flexible income processes in large admin datasets, this paper turns to VI for computational feasibility.
- However, in seeking to make the most of VI, they forgo ABB models and instead explore a class of location-scale models (LSM), which is of independent interest.
- Nevertheless, it would be good to know if we can go beyond LSM and still be able to take advantage of VI.
- The PSID results based on LSMs are interesting, especially the finding that their estimates can mimic nonlinear persistence in ABB.
- The ABB persistence measure is $\rho(z,u) = \partial Q(z,u)/\partial z$ where $Q(z,u)$ is the quantile function of z_t conditioned on $z_{t-1} = z$, and their finding is:

$$\begin{aligned} \rho\left(z^{rich}, 0.9\right) - \rho\left(z^{rich}, 0.1\right) &> 0 \\ \rho\left(z^{poor}, 0.9\right) - \rho\left(z^{poor}, 0.1\right) &< 0 \end{aligned}$$

- In the LSM:

$$\rho(z, 0.9) - \rho(z, 0.1) = \sigma'(z)\kappa \qquad \kappa > 0.$$

- Therefore, it can mimic the ABB finding as long as $\sigma'\left(z^{rich}\right) > 0$ and $\sigma'\left(z^{poor}\right) < 0$, which seems consistent with the tilted U-shape BBL find for $\sigma(z)$.

Location-scale models (continued)

- Another key finding in ABB concerns conditional asymmetry:

$$sk\left(z^{rich}\right) < 0, \qquad sk\left(z^{poor}\right) > 0$$

  where $sk(z)$ is Kelley's skewness measure.

- In a Gaussian LSM $sk(z) = 0$. In a skewed LSM, $sk(z) \neq 0$, but cannot vary with z.

- Interestingly, BBL report that discounted densities exhibit ABB-type skewness.

- This is plausible because, while $z_t \mid z_{t-1}$ follows a LSM, the two-step-ahead distribution $z_t \mid z_{t-2}$ no longer has an LS form but instead becomes a normal-mixture.

- Given the nonlinear persistence induced by $\sigma(z)$, it is not entirely surprising that discounted densities exhibit ABB asymmetries, but the way they do so is tied to $\sigma(z)$.

- Since there is no free third-order moment $sk(z)$ that can vary independently with z, in the discounted densities both dispersion and skewness are driven by $\sigma(z)$.

- To what extent this is restrictive is an interesting question to elucidate.

Policy-relevant quantities as diagnostics

- The comparisons between true and approximate costs of risk and certainty equivalents provide useful diagnostics based on metrics with an economic interpretation.
- Beyond this, it would be interesting to examine some policy-relevant quantities.
- For instance, Guner, Renedo, and Yavuz (2025) show that under an ABB process, heterogeneity in the value of social insurance increases along the income distribution.
- They find that social insurance programs are less valuable for poorer households and more valuable for richer households.
- As another example, De Nardi, Fella, and Paz-Pardo (2025) analyze optimal benefit configurations—the balance between income support and in-work benefits.
- With an ABB wage process, the optimal design favors income support as the main component of benefits, regardless of labor force participation.
- By contrast, under a canonical wage process, the optimal mix tilts toward providing work incentives rather than insurance against low labor income outcomes.

Modelling options

- Since the motivation is to estimate income processes using big data, it is worth reviewing modeling choices.
- With richer admin data, we may not wish to rely on the same types of models that have traditionally been estimated on survey data.
- I discuss some related aspects next.

Why include transitory shocks?

- This is relevant to ask because the distinction between permanent and transitory is the reason why estimating these models is challenging with large databases.
- The usual justification is to recognize that income can be affected by different types of shocks. Some shocks have more persistent effects on income than others, and these differences lead to different effects on consumption behavior.
- The p-t formulation is the poor-person option to have more than one income shock when only income data are available.
- By labeling one component as persistent and the other as transitory, we can have two shocks from only one observable—an idea that goes back to Friedman and Kuznets.
- However, big admin data not only provide information on the income of more people, but often also offer access to observable shocks of different nature or frequency.
- To be sure, panel surveys have long combined individual incomes with employment histories, which themselves are natural sources of multiple shocks affecting income.
- Multivariate approaches have yielded important contributions (Low, Meghir & Pistaferri 2010; Altonji, Smith & Vidangos 2013; Friedrich, Laun, Meghir &Pistaferri 2019), but scaling this work effectively remains a challenge.

Unobserved heterogeneity

- Flexible modelling of income and consumption dynamics has tended to downplay unobserved heterogeneity and the current paper is not an exception.

- Almuzara (2023) shows that linear dependence combined with a heterogeneous transitory shock can mimic nonlinear dynamics, offering a distributional version of the unobserved heterogeneity vs. state dependence dilemma.

- Incorporating unobserved heterogeneity complicates p–t formulations, especially for long panels. For instance, sequential Monte Carlo samplers can no longer be relied upon (Arellano, Blundell, Bonhomme, and Light 2023).

- Likewise, accounting for observed heterogeneity quickly becomes challenging in flexible p–t models.

- Variational inference may provide a promising approach to address both observed and unobserved heterogeneity in flexible dynamic models.

High frequency vs low frequency

- In Social Security admin data, we often observe high-frequency earnings information alongside real-time work histories.

- High-frequency and low-frequency shocks affect income dynamics differently.

- High-frequency earnings data can be decomposed into latent factors that operate at different frequencies, e.g. as in the approach of Forni, Hallin, Lippi & Reichlin (2000).

- Similarly, ARMA models with seasonal components–such as monthly and annual cycles–can be equivalently represented using dynamic factor models.

- The idea that latent factors at various frequencies can be modeled as location-scale processes, and that their aggregation yields flexible representations of earnings dynamics, is closely aligned with the BBL framework.

Long panels or long time series of panels?

- Sometimes, two distinct time dimensions play very different roles from an identification perspective.

- The first is the number of observations over time per unit, denoted S. A value of $S > 1$ is essential for identifying transitions and allowing for unobserved heterogeneity.

- However, even a relatively small S can suffice to identify Markovian dynamics with low-dimensional unobserved heterogeneity.

- The second is the calendar time spanned by the units in the dataset, denoted T, which is key to capture business cycle fluctuations and policy regime changes.

- The T dimension is critical for identifying aggregate effects and interactions between aggregate and idiosyncratic sources of variation (Almuzara et al 2025).

- The distinction between S and T also matters computationally: long panels tracking the same individuals are more demanding than long time series of short panels.

Concluding remarks

- BBL began by expressing interest in evaluating the usefulness of variational methods for flexible income processes, with the goal of applying them to large datasets.

- So far they focused on a class of models that, at first glance, appear less flexible than those typically emphasized in the literature, and estimated them using PSID data.

- Nonetheless, their results are very interesting in their own right.

- An open question remains: how far can these models and their associated variational posteriors be pushed before sacrificing the computational advantages of VI?

- Another important question is which classes of models and variational approximations yield consistent estimators of the underlying likelihood parameters.

- Finally, as we move toward large-scale data, it may be time to revisit the models we target—regarding persistent vs transitory components, observed vs unobserved heterogeneity, high- vs low-frequency, and the time structure of long panels.