How Variational Are Earnings Dynamics?*

Neele Balke
University of Chicago

Stéphane Bonhomme
University of Chicago

Thibaut Lamadon
University of Chicago

December 6, 2025

Abstract

Variational inference is a popular machine learning method in a variety of fields, including natural language processing and computer vision. In this paper we evaluate its performance to estimate models of earnings dynamics. We focus on models with a Markovian, conditionally Gaussian persistent component and an additive transitory component. We parameterize the data-generating processes to match estimates from the earnings literature while allowing for nonlinearities, non-Gaussianity, serial correlation in transitory shocks, and time-invariant latent heterogeneity. We implement a range of variational posterior families leading to differentiable objective functions. The results highlight that variational posterior choice is crucial: independent (also known as "mean-field") approximations systematically underperform, while richer families provide more reliable inference. We find that the persistent component process is generally well recovered, but that the kurtosis of transitory shocks tends to be understated. Finally, applying the estimator to the Panel Study of Income Dynamics (PSID) from 1980 to 1990, we find the presence of nonlinearities in the conditional volatility and persistence of earnings.

JEL codes: C10. C50.

Keywords: Earnings dynamics, Variational inference, Posterior distributions, Nonlinear state-space models.

*We thank our discussant Manuel Arellano, as well as the audience of the 2025 Yale conference in honor of Costas Meghir, for helpful comments. This research was supported in part by the Pythia computing cluster at the University of Chicago Booth School of Business, which is funded by the Office of the Dean.

1 Introduction

Understanding the nature of individual earnings dynamics is central to answering a wide range of economic questions, from the design of social insurance systems to the evaluation of consumption and saving behavior. While much empirical work has relied on linear specifications of earnings processes, growing evidence (e.g., Guvenen et al., 2021, Arellano et al., 2017) suggests that nonlinearities – such as asymmetric shocks and nonlinear persistence – play a crucial role in shaping economic outcomes; see the survey by Arellano (2014). What is the shape of these nonlinearities? How much risk do individuals face at different points of the income distribution? How much of this risk is transitory, how much is persistent, and how much is reflected into consumption decisions?

Answering these questions requires empirical frameworks that go beyond linear autoregressions, motivating dynamic, nonlinear state-space models with latent variables – such as the model proposed by Arellano et al. (2017) – to flexibly capture the joint dynamics of earnings through the dynamics of their persistent and transitory components. However, despite their appeal, such nonlinear latent variable models are computationally demanding to estimate. The core challenge comes from the difficulty to evaluate the likelihood function. In models with transitory shocks, the likelihood function does not admit a closed-form expression and must instead be approximated numerically. Available strategies in the literature, such as based on extensions of the Expectation–Maximization (EM) algorithm, require sophisticated methods to draw the latent components, which are delicate to tune and do not scale well to longer panels.

This paper analyses a scalable estimation approach for nonlinear latent variable models, leveraging recent advances in machine learning. We employ a *variational inference* approach (e.g., Jordan et al., 1999, Kingma and Welling, 2014), which maximizes the so-called Evidence Lower Bound (ELBO) instead of the log-likelihood function; see Blei et al. (2017) for a comprehensive survey. Variational inference transforms the problem of repeated integration over latent trajectories into an optimization problem, thereby avoiding the curse of dimensionality that undermines exact likelihood-based methods. It is fully differentiable and naturally compatible with modern automatic differentiation and stochastic gradient algorithms, which makes estimation fast and scalable to panels with long time series or large populations – even when the latent process is nonlinear or non-Gaussian. These features make it feasible to bring flexible state-space models to data that were previously beyond reach.

At the same time, the approach comes with important limitations. Because the ELBO is only a lower bound on the log-likelihood function, maximizing the ELBO generally leads to

a biased and inconsistent estimator. The ELBO objective is constructed by approximating the posterior distribution of the latent components (which is intractable in the setting we study) by a family of *variational posterior densities*. The bias on parameter estimates tends to increase with the gap between the true posterior and the variational family. Moreover, while considering a flexible family reduces this gap, the optimization problem can be complex, requiring careful parameterization of the posterior and attention to convergence issues. In practice, the effectiveness of variational inference thus depends critically on the design of the variational posterior family: too restrictive a form risks incorrect inference, while overly flexible forms may be computationally burdensome or unstable.

Our starting point is a nonlinear latent variable model of earnings dynamics that nests as a special case the "canonical" persistent-transitory model: a linear Gaussian autoregressive process with an additive i.i.d. shock. The generalization we introduce accommodates nonlinear persistence, state-dependent volatility, serially correlated or heavy-tailed transitory shocks, and time-invariant latent heterogeneity. In the model, the persistent component evolves through general conditional mean and volatility, while transitory shocks follow a flexible distribution that may exhibit excess kurtosis, moving average dependence, or heterogeneous volatility across individuals. This model captures many of the salient features emphasized in the empirical literature, including state-dependent volatility, asymmetric shocks, and nonlinear persistence.

We use simulated data to evaluate how variational inference performs across a sequence of models of increasing complexity: a linear Gaussian benchmark, a specification where mean and variance are nonlinear functions of the persistent state, a model where transitory shocks follow a first-order moving average process, and a model where the variances of transitory shocks differ across individuals. In all cases, we rely on a Gaussian specification to approximate the posterior distribution of the latent components. While the choice of the Gaussian is motivated by its simplicity, we experiment with various restrictions on the covariance matrix, and in extensions we also evaluate the performance of some specific transformations of the Gaussian.

We start by considering the canonical model based on a linear Gaussian process. We find that the parameters of this benchmark model are well recovered. This is expected since in this case the true posterior is Gaussian, and thus belongs to the variational family we rely on. However, we find that correctly capturing the covariance structure of the earnings process is crucial: when using a variational posterior with independent components over time (a popular specification in machine learning called "mean-field" approximation) biases are substantial.

We then consider nonlinear extensions of this model. We find that variational inference recovers the conditional mean and variance of the persistent component quite well. However,

it does not capture higher moments of the transitory component such as kurtosis. Similar difficulties arise in versions of the model that include serially correlated transitory shocks or time-invariant heterogeneity. Across specifications, we find that mean-field approximations perform poorly, while unrestricted Gaussian variational posteriors and their structured counterparts that exploit the dynamic structure of the model perform better.

We apply the method to annual data from the Panel Study of Income Dynamics (PSID) for the years 1980–1990, estimating a flexible model that allows for an MA(1) transitory component and nonlinearity in the process of the persistent component. We find that the conditional mean of the persistent earnings component is approximately linear, with average persistence close to unity, while the conditional variance is nonlinear, displaying a U-shape across the distribution. We also uncover evidence of serial correlation in transitory shocks. As in Arellano et al. (2017), we find that persistence is nonlinear, in the sense that it is lower when high-income households experience negative shocks or low-income households experience positive shocks.

Overall, we see this paper as a primer on the use of variational inference for estimating nonlinear models of earnings dynamics. We restrict ourselves to a Gaussian variational family, or (in an extension) to a simple transformation of the Gaussian. Despite this parsimonious choice, the evidence we obtain suggests that, while the method tends to understate the non-Gaussian features of transitory shocks, in the settings we study the variational approximation captures some of the key features of persistence, volatility, and risk that matter for economic analysis. This encouraging evidence motivates importing and improving these methods for their use in dynamic economic settings.

Literature. A rich empirical literature has studied the dynamics of individual earnings, documenting important linear and nonlinear dynamic features. Key contributions include Lillard and Willis (1978) and Abowd and Card (1989), among many others. Early work modeled earnings as the sum of persistent and transitory components, typically assuming linear Gaussian processes. However, subsequent empirical studies have shown that these assumptions are too restrictive. Meghir and Pistaferri (2004) develop econometric methods to separate transitory and permanent shocks to income using both earnings and consumption data, uncovering heterogeneity in variances across individuals. The broader literature, as reviewed by Meghir and Pistaferri (2011), emphasizes the importance of nonlinearities, including age-dependent volatility, state-dependent persistence, and heteroskedasticity as a source of earnings risk.

Recent contributions use nonlinear state-space models to allow for richer dynamics. Arellano et al. (2017) propose a quantile-based panel framework in which individual latent states

evolve nonlinearly. This setup reveals that persistence is nonlinear, showing that shocks have different effects depending on past earnings. De Nardi et al. (2020) compare linear and nonlinear earnings processes in a structural life-cycle model and show that models with skewness, kurtosis, and state-dependent variances produce markedly different predictions for consumption inequality and self-insurance behavior than their linear Gaussian counterparts. Braxton et al. (2024) develop a generalized Kalman filter and document how income risk varies along the skill distribution.

While variational inference has been successfully applied to many fields, including text analysis (Blei et al., 2003) and computer vision (Kingma and Welling, 2014), applications to economics are still relatively limited. Chan and Yu (2022) develop variational methods for large Bayesian VARs with stochastic volatility. Loaiza-Maya and Nibbering (2023) apply variational inference to structural discrete choice models. Mele and Zhu (2023) and Bonhomme (2021) use variational inference to estimate network formation and team production models, respectively.

In this paper, we focus on the empirical performance of variational inference on simulated and real data. There is also a literature on theoretical properties. A key question is whether parameters that maximize the evidence lower bound are asymptotically consistent for the true parameters of the model. Bickel et al. (2013) provide results for stochastic blockmodels, Westling and McCormick (2019) study Gaussian mixture models, and Katsevich and Rigollet (2024) study Gaussian variational inference when the true posterior is approximately Gaussian. In recent work, Medina et al. (2022) study the use of α-posteriors and their variational approximations in the presence of model misspecification.

The remainder of the paper is organized as follows. In Section 2 we introduce a nonlinear latent variable model of earnings dynamics, and in Section 3 we describe how variational inference can serve as a practical alternative to direct likelihood methods in this context. We then present simulation evidence based on a linear Gaussian data-generating process (DGP) in Section 4, and based on a nonlinear DGP in Section 5. In Section 6 we show how to extend the approach to allow for serially correlated transitory shocks and time-invariant latent heterogeneity. In Section 7 we present an empirical application using PSID data. Finally, in Section 8 we describe two alternatives to Gaussian-based variational inference, and we conclude in Section 9. Details about implementation are provided in Appendix B, and computer codes will be available online.

2 A Nonlinear Model of Earnings Dynamics

In this section, we introduce a nonlinear latent variable model for individual earnings dynamics. The model has a hidden Markov – i.e., permanent-transitory – structure. Let y_t denote log earnings for an individual at time t, and let z_t denote a latent component that evolves according to a state-dependent stochastic process. The observed data, for a given individual, are generated by:

$$y_t = z_t + e_t, \tag{1}$$

$$z_t = \mu(z_{t-1}) + \sigma(z_{t-1})u_t, \tag{2}$$

$$z_1 \sim f_\alpha, \quad u_t \sim \mathcal{N}(0,1), \quad e_t \sim \psi_\gamma, \tag{3}$$

where u_t and e_t are independent at all lags and independent of the initial z_1, and periods range from 1 to T.

In this formulation, z_t represents the persistent component of individual earnings, while e_t are transitory shocks. The innovations u_t are standard normal, scaled by a state-dependent volatility function $\sigma(z_{t-1})$, and propagated forward through a conditional mean function $\mu(z_{t-1})$. Both μ and σ are specified flexibly, either as low-order polynomials or neural networks, so that the model nests the standard linear earnings process, but also allows for nonlinear and state-dependent dynamics. The initial latent state z_1 is drawn from a distribution f_α, which can itself be parameterized using flexible location-scale families to capture cross-sectional heterogeneity in initial conditions.

The distribution of transitory shocks, ψ_γ, is modeled flexibly. Alongside a Gaussian benchmark case, we allow for non-Gaussian specifications that capture excess kurtosis, as well as specifications that introduce serial dependence through an MA(1) process. This flexibility enables the model to accommodate a wide range of transitory earnings innovations, including asymmetry and thick tails, which are commonly observed in administrative and survey data.

This structure nests an array of familiar models as special cases. Standard linear AR(1) models with Gaussian shocks are obtained when the conditional mean $\mu(z_{t-1})$ is linear, the volatility function $\sigma(z_{t-1})$ is constant, and the transitory component ψ_γ is jointly normally distributed. At the same time, the specification is rich enough to encompass more complex earnings processes that have been emphasized in recent empirical work. For example, Arellano et al. (2017) highlight the importance of nonlinear persistence and state-dependent volatility, while Meghir and Pistaferri (2004) provide evidence of heterogeneity in both the magnitude and persistence of permanent and transitory shocks. Our framework is designed to capture precisely

these features, allowing for greater flexibility while nesting a number of existing models. In addition, we will show in Section 6 that it can be augmented to allow for the presence of latent time-invariant heterogeneity.

Remark 1. *(Conditional skewness) Model (1)-(2)-(3) imposes that z_t is Gaussian given z_{t-1}, thus ruling out non-Gaussian features in the conditional distribution such as conditional skewness (Arellano et al., 2017). However, note that, even when both u_t and e_t are symmetrically distributed, the combination of a nonlinear drift and conditional volatility in the persistent component introduces higher-order dependence and conditional asymmetries in the distribution of earnings changes. In particular, the model can generate conditional skewness and excess kurtosis over multiple periods ahead, even when one-period-ahead shocks are Gaussian. Our empirical application in Section 7 will illustrate this.*

3 Variational Inference

The model given by equations (1)-(2)-(3) is parametric, indexed by the parameters $\mu(\cdot)$, $\sigma(\cdot)$, α, and γ. For conciseness we will use $\theta = (\mu(\cdot), \sigma(\cdot), \alpha)$ to denote all the parameters that index the distribution of the persistent component z_t, whereas γ indexes the distribution of transitory shocks e_t. The econometrician seeks to learn these parameters based on a sequence of outcomes $y_1, ..., y_T$, available for a collection of individuals, although we omit the individual dimension from the notation for simplicity.

3.1 Issues with Evaluating the Likelihood

The log-likelihood of the observed data for a given individual can be expressed as

$$\mathcal{L}_{\theta,\gamma}(y_{1:T}) = \log \mathcal{P}_{\theta,\gamma}(y_{1:T}) = \log \int f_\theta(z_{1:T})\, \psi_\gamma(y_{1:T} - z_{1:T})\, dz_{1:T}, \tag{4}$$

where $f_\theta(z_{1:T})$ denotes the density of $z_{1:T} = (z_1, ..., z_T)$ over latent trajectories parameterized by θ, and $\psi_\gamma(y_{1:T} - z_{1:T})$ is the conditional density of outcomes given $z_{1:T}$, parameterized by γ, linking latent states to observed data. Equation (4) shows that evaluating the marginal log-likelihood $\mathcal{L}_{\theta,\gamma}(y_{1:T})$ requires integrating over all possible latent trajectories. However, this task is generally infeasible in realistic models of earnings dynamics.[1]

[1]In this paper, we view the log-likelihood function as the target objective to optimize. This presumes that the conditions for consistency of maximum likelihood are satisfied. An important necessary condition is identification, which may be challenging to establish in these models, see among others Hu and Schennach (2008) and Arellano et al. (2017).

To see the difficulty, consider evaluating (4) for one individual. Since there are T latent variables, $z_1, ..., z_T$, numerical quadrature or grid-based methods become inaccurate as soon as there are more than a handful of periods. A common approach is to resort to simulation-based techniques such as importance sampling, particle filtering, or Markov Chain Monte Carlo, see for example Creal (2012) for a survey of particle filter methods and Arellano et al. (2024) for an application in the context of earnings and consumption dynamics. However, while such approaches perform well in the case of a single time series, evaluating as many integrals in (4) as there are individuals in the sample represents a formidable challenge. As a result, state-of-the-art algorithms for nonlinear models of earnings dynamics based on these techniques are currently limited to data sets of moderate dimension, especially regarding the number of time periods.

3.2 The Evidence Lower Bound

To address the computational challenges of exact likelihood-based methods, we study a *variational inference* approach. Instead of evaluating or maximizing the marginal log-likelihood directly, variational inference reframes the problem as an optimization of the so-called *evidence lower bound* (ELBO):

$$\mathcal{E}_{\theta,\gamma,\phi}(y_{1:T}) = \mathbb{E}_{q_\phi(z_{1:T}\,|\,y_{1:T})}\left[\log \frac{f_\theta(z_{1:T})\psi_\gamma(y_{1:T} - z_{1:T})}{q_\phi(z_{1:T}\,|\,y_{1:T})}\right], \tag{5}$$

where $q_\phi(z_{1:T}\,|\,y_{1:T})$ is some density over the latent states $z_1, ..., z_T$ – the so-called *variational posterior* density – indexed by a parameter vector ϕ, and the expectation is taken with respect to $q_\phi(z_{1:T}\,|\,y_{1:T})$ for a fixed sequence of observations $y_{1:T}$.

To understand the logic behind the maximization of (5), and to see that the ELBO is indeed a lower bound on the log-likelihood function, it is useful to introduce some notation. Interpreting f_θ as a *prior* on the states $z_1, ..., z_T$, denote the associated *posterior* density as, by Bayes' rule,

$$p_{\theta,\gamma}(z_{1:T}\,|\,y_{1:T}) = \frac{f_\theta(z_{1:T})\psi_\gamma(y_{1:T} - z_{1:T})}{\mathcal{P}_{\theta,\gamma}(y_{1:T})}. \tag{6}$$

Note that, since the likelihood function $\mathcal{P}_{\theta,\gamma}(y_{1:T})$ appears in the denominator of (6), the posterior density is typically highly challenging to calculate, for the reasons mentioned in the previous subsection.

Next, let KL[q ‖ p] denote the Kullback–Leibler divergence between q and p; that is,

$$\mathrm{KL}[q \,\|\, p] = \int \log\left(\frac{q(z)}{p(z)}\right) q(z)dz.$$

Observe that the ELBO can equivalently be written as

$$\begin{aligned}\mathcal{E}_{\theta,\gamma,\phi}(y_{1:T}) &= \mathbb{E}_{q_\phi(z_{1:T}\,|\,y_{1:T})}\left[\log \frac{f_\theta(z_{1:T})\psi_\gamma(y_{1:T}-z_{1:T})}{q_\phi(z_{1:T}\,|\,y_{1:T})}\right] \\ &= \log \mathcal{P}_{\theta,\gamma}(y_{1:T}) - \mathbb{E}_{q_\phi(z_{1:T}\,|\,y_{1:T})}\left[\log \frac{q_\phi(z_{1:T}\,|\,y_{1:T})}{\frac{f_\theta(z_{1:T})\psi_\gamma(y_{1:T}-z_{1:T})}{\mathcal{P}_{\theta,\gamma}(y_{1:T})}}\right],\end{aligned}$$

where we have used that $\mathcal{P}_{\theta,\gamma}(y_{1:T})$ does not depend on the latent states $z_{1:T}$. Hence, using the expressions of the posterior density and KL divergence, we obtain the following key identity:

$$\mathcal{E}_{\theta,\gamma,\phi}(y_{1:T}) = \mathcal{L}_{\theta,\gamma}(y_{1:T}) - \mathrm{KL}\big[q_\phi(z_{1:T}\,|\,y_{1:T})\,\|\,p_{\theta,\gamma}(z_{1:T}\,|\,y_{1:T})\big]. \tag{7}$$

Equation (7) shows that the ELBO is equal to the log-likelihood function minus a penalty term that is equal to the KL divergence between the variational posterior density q_ϕ and the true posterior density. This characterization has several important implications, the first one being, since the KL divergence is non-negative, that the ELBO is indeed a lower bound on the log-likelihood.

A second implication of (7) is that maximizing the ELBO with respect to ϕ, for given parameters θ and γ, is equivalent to finding the variational posterior q_ϕ that is closest, in a KL sense, to the true posterior evaluated at θ, γ. When the variational family q_ϕ is very flexible (e.g., when ϕ is high-dimensional), one expects the resulting KL term to be small, and the ELBO and log-likelihood to be close to each other. Indeed, in the case where the variational family includes the true posterior, ELBO and log-likelihood coincide. However, for a restricted variational family such as Gaussian densities, the difference between the log-likelihood and the ELBO (the so-called "ELBO gap") may be substantial.

A third implication of (7) is computational. Note that the likelihood function $\mathcal{P}_{\theta,\gamma}(y_{1:T})$, which is an intractable integral, does not appear in (5). By replacing the log of an expectation with an expectation of logs in (5), relying on the ELBO instead of the log-likelihood transforms the problem into one that is computationally tractable and yields stable gradient estimates. Derivatives of the objective function can now be computed and averaged over, rather than requiring integration over the full latent space. This can provide important computational advantages compared to traditional methods such as the EM algorithm, as we now illustrate.

3.3 The EM Algorithm: Alternating Optimization

A common strategy in latent variable models is to rely on the Expectation–Maximization (EM) algorithm. To relate the latter to the ELBO, suppose that we take $\phi = (\theta', \gamma')$, for some

hypothetical values of the parameters, and set q_ϕ to be the true posterior at those parameters; that is,

$$q_\phi(z_{1:T} \mid y_{1:T}) = p_{\theta',\gamma'}(z_{1:T} \mid y_{1:T}). \tag{8}$$

By (7) we have

$$\mathcal{E}_{\theta,\gamma,\theta',\gamma'}(y_{1:T}) = \mathcal{L}_{\theta,\gamma}(y_{1:T}) - \mathrm{KL}\big[p_{\theta',\gamma'}(z_{1:T} \mid y_{1:T}) \,\|\, p_{\theta,\gamma}(z_{1:T} \mid y_{1:T})\big]. \tag{9}$$

The EM algorithm maximizes $\mathcal{E}_{\theta,\gamma,\theta',\gamma'}(y_{1:T})$, by alternating between two steps:

- **Maximize with respect to θ', γ' (E-step).**

  Given θ, γ, we see from (9) that the maximum with respect to θ', γ' is achieved when

  $$p_{\theta',\gamma'}(z_{1:T} \mid y_{1:T}) = p_{\theta,\gamma}(z_{1:T} \mid y_{1:T}).$$

- **Maximize with respect to θ, γ (M-step).**

  Given θ', γ', the bound (9) attains its maximum when θ, γ maximize

  $$\mathbb{E}_{p_{\theta',\gamma'}(z_{1:T} \mid y_{1:T})} \left[\log f_\theta(z_{1:T}) \psi_\gamma(y_{1:T} - z_{1:T})\right].$$

Notice that the ELBO, and hence the log-likelihood, are weakly increasing in each (E,M) iteration. The EM algorithm can thus be used as an alternative to gradient-based maximization of the likelihood (Dempster et al., 1977). However, the E-step requires evaluating the exact posterior $p_{\theta',\gamma'}(z_{1:T} \mid y_{1:T})$, which is generally intractable. Sampling-based approximations (using, e.g., MCMC or particle filter methods) can be employed, but they are computationally costly, difficult to differentiate, and not easily integrated into optimization routines that rely on automatic differentiation and gradient-based updates.

3.4 Variational Inference: A Fully Differentiable Approach

Variational inference relies on a different approach. The key idea is to replace the exact posterior in (8) with a parameterized approximating family $q_\phi \in Q_\phi$. Instead of computing the true posterior in the E-step, we optimize the variational parameters ϕ to minimize the KL divergence to the (unknown) true posterior. This is equivalent to maximizing the ELBO:

$$\max_{\theta,\gamma,\phi} \mathcal{E}_{\theta,\gamma,\phi}(y_{1:T}) = \max_{\theta,\gamma,\phi} \mathbb{E}_{q_\phi(z_{1:T} \mid y_{1:T})} \left[\log \frac{f_\theta(z_{1:T}) \psi_\gamma(y_{1:T} - z_{1:T})}{q_\phi(z_{1:T} \mid y_{1:T})}\right]. \tag{10}$$

A challenge is then to compute gradients of the ELBO with respect to the variational parameters ϕ, since the expectation $\mathbb{E}_{q_\phi(z_{1:T} \mid y_{1:T})}[\cdot]$ depends on ϕ, preventing the use of standard gradient-based optimization.

The Reparameterization Trick: Enabling Gradient-Based Optimization. The *reparameterization trick* resolves this difficulty by rewriting random draws from q_ϕ as deterministic transformations of the variational parameters, ϕ, and a vector of auxiliary noise random variables, v:

$$z_{1:T} = g(\phi, v; y_{1:T}), \quad v \sim \pi(v).$$

For example, if $q_\phi(z_{1:T} \mid y_{1:T})$ is a Gaussian density with mean μ_ϕ and variance Σ_ϕ, then $z_{1:T} = \mu_\phi + L_\phi v$ with $v \sim \mathcal{N}(0, I)$ and $L_\phi L_\phi^\top = \Sigma_\phi$.[2]

This transformation allows gradients to pass through the expectation, since:

$$\nabla_\phi \mathcal{E}_{\theta,\gamma,\phi}(y_{1:T}) = \mathbb{E}_{\pi(v)}\left[\nabla_\phi \left(\log \frac{f_\theta(g(\phi, v; y_{1:T}))\psi_\gamma(y_{1:T} - g(\phi, v; y_{1:T}))}{q_\phi(g(\phi, v; y_{1:T}) \mid y_{1:T})}\right)\right]. \tag{11}$$

This enables Monte Carlo gradient estimates, and makes the variational objective amenable to automatic differentiation. The reparameterization trick thus provides a fully differentiable alternative to sampling-based methods, allowing variational inference to be integrated with modern deep learning frameworks and deterministic or stochastic gradient optimization.

Amortization. In large datasets, it may be challenging to optimize separate variational parameters ϕ for each observation sequence $y_1, ..., y_T$. *Amortized* variational inference relies on a model

$$\phi = h_\eta(y_{1:T}), \tag{12}$$

where the global parameter η is shared across observation sequences $y_{1:T}$. Model (12) is referred to as a *recognition model* or *encoder* in the literature. In applications, h_η is typically specified as a neural network or other flexible function. Arguments often provided in favor of amortization are that it shares statistical strength across observations, reduces the number of free parameters to be estimated, and enables fast computation by evaluating $h_\eta(y_{1:T})$ rather than re-optimizing a new ϕ for each observation sequence. Amortization is a key feature of variational autoencoders (Kingma and Welling, 2014), a leading application of variational inference.

A Quasi-Likelihood Interpretation. Variational inference relies on maximizing the ELBO. Equivalently, this can be interpreted as maximizing the following *quasi-log-likelihood*:

$$\begin{aligned}\mathcal{E}^*_{\theta,\gamma}(y_{1:T}) &= \max_\phi \mathcal{E}_{\theta,\gamma,\phi}(y_{1:T}) \\ &= \mathcal{L}_{\theta,\gamma}(y_{1:T}) - \min_{q_\phi \in \mathcal{Q}_\phi} \mathrm{KL}\big[q_\phi(z_{1:T} \mid y_{1:T}) \,\|\, p_{\theta,\gamma}(z_{1:T} \mid y_{1:T})\big].\end{aligned} \tag{13}$$

[2]Note that both μ_ϕ and Σ_ϕ may, and typically will, depend on the observation sequence $y_1, ..., y_T$.

Estimates of parameters θ, γ are then obtained as

$$(\widehat{\theta}, \widehat{\gamma}) = \max_{\theta,\gamma} \mathcal{E}^*_{\theta,\gamma}(y_{1:T}).$$

The second term on the right-hand side of (13) acts as a penalty, which distorts the variational inference estimates away from the maximum likelihood estimator. In particular, unlike the expected log-likelihood, the expected quasi-log-likelihood $\overline{\mathcal{E}}(\theta, \gamma) = \mathbb{E}[\mathcal{E}^*_{\theta,\gamma}(y_{1:T})]$ is not maximized at true parameter values in general, and variational estimators do not converge to true parameter values in large samples. This issue is the main challenge in applying and interpreting variational inference, and our goal in the rest of the paper is to evaluate the performance of the method on simulated and empirical data.

4 A Linear Gaussian Model with Closed-Form Likelihood and Posterior

To benchmark our variational approximations and evaluate their accuracy in a controlled setting, in this section we consider a simple linear Gaussian specification of the latent variable model introduced in Section 2. In this version, the persistent and transitory components follow linear Gaussian processes, allowing for closed-form expressions of the likelihood and posterior distribution. This setup serves as a benchmark for evaluating variational inference on nonlinear models in later sections.

4.1 The Linear Gaussian Model

We consider the following specification of the data-generating process:

$$y_t = z_t + e_t, \tag{14}$$

$$z_t = \rho z_{t-1} + \sigma u_t, \tag{15}$$

$$z_1 \sim \mathcal{N}(0, \sigma^2_{z_1}), \quad u_t \sim \mathcal{N}(0,1), \quad e_t \sim \mathcal{N}(0, \sigma^2_e). \tag{16}$$

Here, z_t is an unobserved latent state evolving as an AR(1) process with innovation variance σ^2, and y_t is the observed outcome (e.g., log-earnings), measured with additive Gaussian transitory shocks of variance σ^2_e. The initial condition is assumed to be normally distributed, $z_1 \sim \mathcal{N}(0, \sigma^2_{z_1})$, which allows for non-stationary initial conditions. Observations are i.i.d. across individuals (although we again omit the individual subscript from the notation for simplicity). To map this simple model to the notation of the earlier sections, here θ includes ρ, σ, and σ_{z_1}, and $\gamma = \sigma_e$.

Closed-Form Likelihood. Since the model is linear and Gaussian, the joint distribution of latent states and observations is multivariate Gaussian. Specifically, we have

$$z_{1:T} \sim \mathcal{N}(0, \Sigma_z), \tag{17}$$

$$y_{1:T} \,|\, z_{1:T} \sim \mathcal{N}(z_{1:T}, \sigma_e^2 I_T), \tag{18}$$

where

$$\Sigma_z = \begin{pmatrix} V_1 & \rho V_1 & \rho^2 V_1 & \cdots & \rho^{T-1} V_1 \\ \rho V_1 & V_2 & \rho V_2 & \cdots & \rho^{T-2} V_2 \\ \rho^2 V_1 & \rho V_2 & V_3 & \cdots & \rho^{T-3} V_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} V_1 & \rho^{T-2} V_2 & \rho^{T-3} V_3 & \cdots & V_T \end{pmatrix},$$

for (assuming $|\rho| < 1$)

$$V_t = \rho^{2(t-1)} \sigma_{z_1}^2 + \sigma^2 \frac{1 - \rho^{2(t-1)}}{1 - \rho^2}, \qquad t = 1, \ldots, T.$$

The marginal likelihood of the data is then obtained by integrating out the latent variables, which gives

$$y_{1:T} \sim \mathcal{N}(0, \Sigma_z + \sigma_e^2 I_T). \tag{19}$$

Therefore, the log-likelihood function is:

$$\mathcal{L}_{\rho,\sigma,\sigma_{z_1},\sigma_e}(y_{1:T}) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_z + \sigma_e^2 I_T| - \frac{1}{2} y_{1:T}^\top (\Sigma_z + \sigma_e^2 I_T)^{-1} y_{1:T}. \tag{20}$$

This expression can be evaluated exactly, as Σ_z^{-1} has a known tridiagonal structure that reflects the AR(1) dynamics. Specifically, we have

$$\Sigma_z^{-1} = \begin{pmatrix} \frac{1}{\sigma_{z_1}^2} + \frac{\rho^2}{\sigma^2} & -\frac{\rho}{\sigma^2} & 0 & \cdots & 0 \\ -\frac{\rho}{\sigma^2} & \frac{1+\rho^2}{\sigma^2} & -\frac{\rho}{\sigma^2} & \cdots & 0 \\ 0 & -\frac{\rho}{\sigma^2} & \frac{1+\rho^2}{\sigma^2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -\frac{\rho}{\sigma^2} \\ 0 & \cdots & 0 & -\frac{\rho}{\sigma^2} & \frac{1}{\sigma^2} \end{pmatrix}. \tag{21}$$

True Posterior Distribution. Since the joint distribution of $(z_{1:T}, y_{1:T})$ is Gaussian, the posterior $p(z_{1:T} \mid y_{1:T})$ is also multivariate Gaussian:

$$z_{1:T} \mid y_{1:T} \sim \mathcal{N}(\mu_{z|y}, \Sigma_{z|y}), \tag{22}$$

where the mean and covariance are given by

$$\mu_{z|y} = \Sigma_z(\Sigma_z + \sigma_e^2 I_T)^{-1} y_{1:T}, \tag{23}$$

$$\Sigma_{z|y} = \Sigma_z - \Sigma_z(\Sigma_z + \sigma_e^2 I_T)^{-1}\Sigma_z, \tag{24}$$

with I_T denoting the $T \times T$ identity matrix. These expressions show that the posterior mean is a linear smoother of the data, and that the posterior covariance shrinks relative to the prior covariance as a function of the signal-to-noise ratio. Note that, while the posterior mean $\mu_{z|y}$ is a linear function of the observed data $y_{1:T}$, the posterior covariance $\Sigma_{z|y}$ depends only on the model parameters and not on the observations sequence.

The posterior covariance matrix $\Sigma_{z|y}$ inherits the Markovian structure of the latent process. Its inverse, the posterior precision matrix $\Omega^{\text{post}} = \Sigma_{z|y}^{-1}$, is obtained as the sum of the prior precision Σ_z^{-1} (which is tridiagonal due to the AR(1) dynamics) and the observation precision $\sigma_e^{-2} I_T$:

$$\Omega^{\text{post}} = \Sigma_z^{-1} + \sigma_e^{-2} I_T. \tag{25}$$

Hence, by (21), the posterior precision matrix remains sparse and tridiagonal, reflecting the fact that each state z_t interacts directly only with its neighbors z_{t-1} and z_{t+1} in the likelihood.

4.2 Variational Posteriors in the Linear Gaussian Model

Implementing variational inference requires defining the variational posterior family q_ϕ to optimize over. We describe here the variational posteriors that we will use in the benchmark model and in the nonlinear frameworks in subsequent sections.

Gaussian Variational Posterior. Since by (22) the true posterior distribution $p(z_{1:T} \mid y_{1:T})$ in the linear-Gaussian model is itself multivariate normal, the Gaussian family is a natural choice for $q_\phi \in \mathcal{Q}_\phi$. That is, for $\mu_q \in \mathbb{R}^T$ and $\Sigma_q \in \mathbb{R}^{T \times T}$ we set the variational parameter as $\phi = (\mu_q, \Sigma_q)$, and the variational posterior density q_ϕ as the $\mathcal{N}(\mu_q, \Sigma_q)$ density:

$$q_\phi(z_{1:T} \mid y_{1:T}) = \frac{1}{(2\pi)^{T/2}\,|\Sigma_q|^{1/2}} \exp\!\left(-\tfrac{1}{2}(z_{1:T} - \mu_q)^\top \Sigma_q^{-1}(z_{1:T} - \mu_q)\right). \tag{26}$$

In the linear-Gaussian model, with the choice (26) of variational density, the evidence lower bound (ELBO) is available in closed form, as

$$\mathcal{E}_{\rho,\sigma,\sigma_{z_1},\sigma_e,\mu_q,\Sigma_q}(y_{1:T}) = -\frac{1}{2}\Big[\sigma_e^{-2}\text{tr}(\Sigma_q + (\mu_q - y_{1:T})(\mu_q - y_{1:T})^\top) + \text{tr}(\Sigma_z^{-1}(\Sigma_q + \mu_q\mu_q^\top)) + \log|\Sigma_z| + T\log\sigma_e^2 - \log|\Sigma_q|\Big] + \text{constant}, \tag{27}$$

where the constant term is irrelevant for optimization. The objective function in (27) is differentiable and can be optimized using standard gradient-based methods.

For implementation, we parameterize the precision matrix Σ_q^{-1} directly using a Cholesky factor, and we compute gradients using the reparameterization trick. Furthermore, we rely on amortized variational inference to specify how both μ_q and Σ_q depend on observations $y_{1:T}$. Following a common practice in the literature (e.g., Kingma and Welling, 2014), we specify the elements of $\mu_q(y_{1:T})$ and the Cholesky factor of $\Sigma_q(y_{1:T})$ as feedforward neural networks.[3] In particular, this implies that the model used in estimation features a very large number of parameters. We will rely on the same approach for nonlinear models in subsequent sections.

A summary of the estimation algorithm is:

1. Initialize μ_q and a Cholesky factor L_q such that $\Sigma_q = L_q L_q^\top$,
2. Sample $z_{1:T} = \mu_q + L_q v$, with $v \sim \mathcal{N}(0, I_T)$,
3. Evaluate the ELBO and its gradient,
4. Update (μ_q, L_q) using (stochastic) gradient descent.

Appendix B provides further details about implementation.

Restricted Gaussian Variational Posteriors. The Gaussian variational family in (26) leaves the parameters μ_q and Σ_q almost unrestricted.[4] However, it may be appealing to impose some of the features of the model's true posterior on the variational family. By reducing the number of parameters involved, restrictions on the variational family may improve estimation accuracy.

We consider three types of restrictions on the Gaussian variational family: under *tridiagonal precision*, a *hidden Markov* structure, and *diagonal precision*, respectively. To motivate the

[3] We first use a shared linear layer with 32 units and a ReLU activation function. The output is then passed to a linear layer without activation to get μ_q, and to a second linear layer with a softplus activation for the elements of Σ_q.

[4] Except for some mild restrictions on their dependence on $y_1, ..., y_T$, modeled using flexible neural networks.

first restriction, note that, by (21), the true posterior in the linear-Gaussian AR(1) benchmark model has a *tridiagonal* precision matrix. It may thus be appealing to impose this restriction on Σ_q^{-1}. Without such a restriction, the variational posterior may introduce spurious dependencies between distant time periods in order to better fit the observed data.

To motivate the second restriction, note that the model has a *hidden Markov* structure, so the true posterior density satisfies:

$$p(z_{1:T} \mid y_{1:T}) = p(z_1 \mid y_{1:T}) \prod_{t=2}^{T} p(z_t \mid z_{t-1}, y_{t:T}),$$

where the last part only depends on future observations $y_t, ..., y_T$. Indeed, conditional on z_{t-1}, past outcomes $y_{1:t-1}$ are redundant in predicting z_t, since their information is mediated entirely through z_{t-1}. This restriction can be incorporated into the variational family as well,

$$q(z_{1:T} \mid y_{1:T}) = q(z_1 \mid y_{1:T}) \prod_{t=2}^{T} q(z_t \mid z_{t-1}, y_{t:T}). \tag{28}$$

In practice, conditioning directly on the entire future sequence $y_{t:T}$ may be computationally burdensome, especially in long panels. As a tractable compromise, one may instead condition on summary statistics of $y_{t:T}$ that preserve essential forward-looking information. Examples include the mean of $y_{t:T}$ or rolling-window averages. These summaries reduce dimensionality while still allowing the variational posterior to capture asymmetry and persistence in the observation sequence.[5]

We will also consider a third type of restriction, imposing that the variational posterior covariance, and hence the posterior precision as well, are *diagonal*. This restriction imposes that all off-diagonal elements of Σ_q^{-1} are equal to zero, hence requiring that, under q_ϕ, the z_t's are independent of each other. While easy to enforce in practice, this restriction does *not* hold under the model, since the first off-diagonal elements of (21) are all non-zero except when $\rho = 0$. Independence assumptions (so-called "mean-field" approximations) are often imposed in applications of variational inference. However, we will see that in models of earnings dynamics they can lead to substantial biases on parameter estimates.

4.3 Results for the Linear Gaussian Model

To evaluate the performance of different variational posterior approximations, we conduct simulation experiments using data generated from Model (14)-(15)-(16). This allows us to directly assess how well different estimators recover the true parameters.

[5]Note that the first two types of restrictions can be combined, and one can jointly enforce that $\Sigma_q^{-1}(y_{1:T})$ is tridiagonal, and that its elements and the elements of $\mu_q(y_{1:T})$ satisfy the hidden-Markov restrictions (28).

Data-Generating Process. In the benchmark design, the latent process z_t follows an AR(1) with persistence parameter $\rho = 0.9$. The initial state has standard deviation $\sigma_{z_1} = 0.4$, innovations are scaled by $\sigma = 0.2$, and the transitory component is normally distributed with standard deviation $\sigma_e = 0.23$. These parameters are chosen to approximately match estimates based on PSID data (e.g., Blundell et al., 2008, Arellano et al., 2017).

Estimation Model and Variational Family. While the true data-generating process is governed by a first-order linear Gaussian model, in the estimation procedure we allow for greater flexibility by specifying the conditional mean and volatility using second-order polynomials:

$$\mu(z_{t-1}) = \mu_0 + \mu_1 z_{t-1} + \mu_2 z_{t-1}^2, \quad \sigma(z_{t-1}) = \log\left(1 + \exp\left(\sigma_0 + \sigma_1 z_{t-1} + \sigma_2 z_{t-1}^2\right)\right).$$

We also specify the distribution of the transitory component e_t and the initial state z_1 to be normal. This setup allows us to assess whether the variational approach can correctly recover the linear DGP by estimating the other polynomial coefficients μ_0, μ_2 and σ_1, σ_2 to be close to zero. Lastly, we compare the results for various choices of variational posterior families, as described above.

Results. Our findings are presented in Figure 1 and Table 1. We find that variational inference based on an unrestricted Gaussian variational posterior performs well in recovering the parameters of the model (first row in Table 1). Imposing a tridiagonal precision matrix leads to very similar estimates (second row).[6] Imposing a hidden Markov structure on the variational posterior, as in (28), shows equivalent performance to the unrestricted and tridiagonal specifications (third row). Hence, imposing restrictions on the variational posterior that are satisfied by the true posterior (at true parameter values) leads to accurate inference in this setting.

However, imposing that Σ_q is diagonal, which is a feature that is *not* present in the true posterior under the DGP, leads to biases (fourth row in Table 1). In particular, the estimate of persistence of z_t is attenuated, with the autoregressive coefficient estimated at 0.75, compared to the true value of 0.9 in the DGP. This downward bias reflects the inability of the diagonal

[6]The choice between the two is a practical trade-off. The unrestricted approach allows for a single Cholesky decomposition of a dense covariance matrix but scales quadratically in the time dimension T, which can become computationally expensive in long panels. In contrast, the tridiagonal specification exploits the sparsity induced by the Markov structure and scales linearly in T, but it typically requires a separate Cholesky decomposition for each observation or variational update step, depending on the implementation. The relative computational efficiency of the two approaches therefore depends on the computational environment, the size of T, and the optimization strategy used.

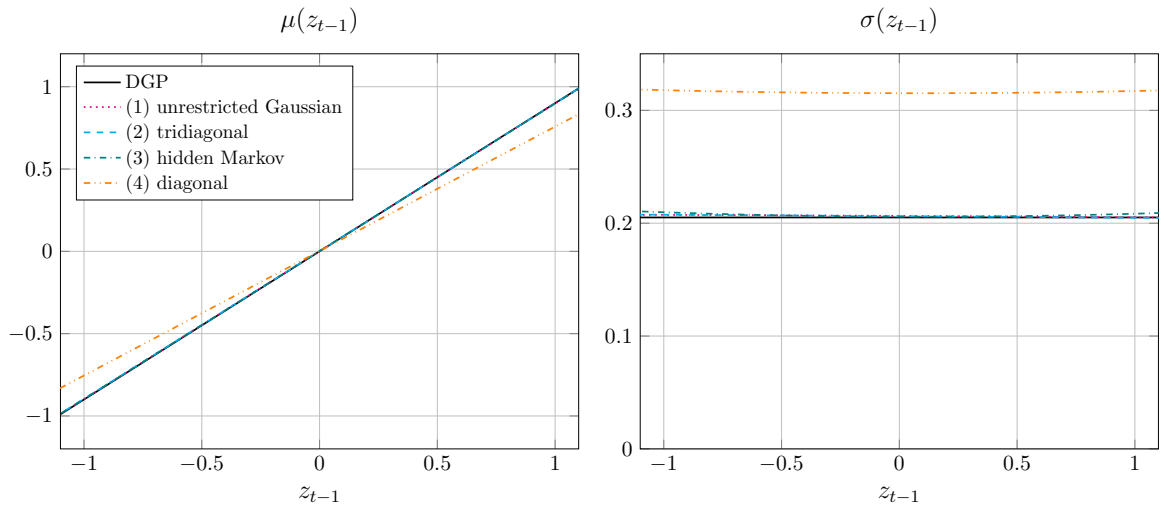

Figure 1: Simulation Results in the Linear Gaussian Model

Note: The figure plots the conditional mean and volatility functions of equation (2) implied by the simulated linear Gaussian DGP, together with their estimated counterparts obtained under four variational posterior specifications. Each panel shows the true functions (solid lines) and the estimated ones (dashed lines) over the support of z_{t-1}.

specification to capture temporal dependencies in the latent states. In addition, the transitory shock standard deviation is underestimated, while the innovation standard deviation of the latent process is overestimated. These distortions indicate a misallocation of uncertainty, as the simplified posterior fails to propagate information across time and compensates by shifting variation across the latent components.

Lastly, to put these findings into perspective, we report in the last row of Table 1 estimates that are based on a Gaussian model without a transitory component. Even in the benchmark AR(1) model, omitting transitory components in observed log-earnings leads to serious estimation bias. Specifically, the estimated persistence parameter is attenuated (as expected in the presence of classical measurement error) and the innovation volatility is systematically underestimated. This underscores the critical importance of accounting for transitory shocks in models of earnings dynamics, as has been extensively demonstrated in the literature.

4.4 Mean-Field Approximation in the Linear Gaussian Model: Analytical Insights

To provide intuition about variational inference in the linear Gaussian model, it is useful to note that, for a Gaussian variational density with parameters $\phi = (\mu_q, \Sigma_q)$, the KL divergence

Table 1: Simulation Results in the Linear Gaussian Model

Parameter	μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	σ_{z_1}	σ_e
DGP	0.00	0.90	0.00	-1.48	0.00	0.00	0.40	0.23
Variational posterior								
(1) unrestricted Gaussian	0.00	0.90	0.00	-1.48	-0.01	0.00	0.40	0.23
(2) tridiagonal	0.00	0.90	0.00	-1.48	-0.01	0.00	0.40	0.23
(3) hidden Markov	0.00	0.90	0.00	-1.48	0.00	0.02	0.40	0.23
(4) diagonal	0.00	0.76	0.00	-1.00	0.00	0.01	0.44	0.13
Ignoring transitory shocks								
(5)	0.00	0.70	0.00	-0.85	-0.01	0.01	0.46	–

Note: The table reports the parameter values used in the simulated DGP and the corresponding estimates obtained under four variational posterior specifications. The estimation model allows for quadratic terms in the conditional mean and log-volatility functions. In the last specification, it abstracts from transitory shocks.

has the following analytical expression:

$$\begin{aligned}
&\mathrm{KL}\big[q_\phi(z_{1:T} \,|\, y_{1:T}) \,\|\, p_{\theta,\gamma}(z_{1:T} \,|\, y_{1:T})\big] \\
&= \frac{1}{2}\left(\log\frac{|\Sigma_{z|y}|}{|\Sigma_q|} - T + \mathrm{tr}\left(\Sigma_{z|y}^{-1}\Sigma_q\right) + \left(\mu_{z|y} - \mu_q\right)^\top \Sigma_{z|y}^{-1}\left(\mu_{z|y} - \mu_q\right)\right),
\end{aligned}$$

where $\mu_{z|y}$ and $\Sigma_{z|y}$ are given by (23) and (24), respectively.

Suppose that μ_q is unrestricted. Further, following the mean-field approach, suppose that Σ_q is diagonal. Then, denoting as $\sigma_{q,t}^2$ the diagonal elements of Σ_q, and using the notation $\Omega^{\text{post}} = \Sigma_{z|y}^{-1}$ for the posterior precision matrix, with diagonal elements $(\omega_t^{\text{post}})^2$, we have

$$\begin{aligned}
&\min_{\mu_q,\Sigma_q} \mathrm{KL}\big[q_\phi(z_{1:T} \,|\, y_{1:T}) \,\|\, p_{\theta,\gamma}(z_{1:T} \,|\, y_{1:T})\big] \\
&= \min_{\sigma_{q,1},\ldots,\sigma_{q,T}} \frac{1}{2}\left(-\log|\Omega^{\text{post}}| - \sum_{t=1}^T \log\sigma_{q,t}^2 - T + \sum_{t=1}^T (\omega_t^{\text{post}})^2\sigma_{q,t}^2\right) \\
&= \frac{1}{2}\left(-\log|\Omega^{\text{post}}| + \sum_{t=1}^T \log(\omega_t^{\text{post}})^2\right) \\
&= \frac{1}{2}\log\frac{|\mathrm{diag}\,\Omega^{\text{post}}|}{|\Omega^{\text{post}}|},
\end{aligned}$$

where $\mathrm{diag}\,\Omega^{\text{post}}$ is the diagonal of Ω^{post}.

Hence, the quasi-log-likelihood in (13) equals

$$\begin{aligned}&\mathcal{E}^*_{\rho,\sigma,\sigma_e,\sigma_{z_1}}(y_{1:T})\\&= \underbrace{-\frac{T}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_z + \sigma_e^2 I_T| - \frac{1}{2}y_{1:T}^\top(\Sigma_z + \sigma_e^2 I_T)^{-1}y_{1:T}}_{\text{log-likelihood}} - \underbrace{\frac{1}{2}\log\frac{|\operatorname{diag}\Omega^{\text{post}}|}{|\Omega^{\text{post}}|}}_{\text{penalty}}.\end{aligned} \tag{29}$$

The penalty term in (29), which is always non-negative, does not depend on the data, only on the parameters. It is a measure of dependence in Ω^{post}, and it is equal to zero when Ω^{post} is diagonal, that is, when the latent types are independent in the true posterior. In the linear Gaussian model analyzed in this section, it follows from (21) and (25) that the penalty is an increasing function of the autoregressive coefficient $|\rho|$. When persistence is high, as is typically observed in earnings data, mean-field approximations based on a diagonal variational posterior covariance matrix distort the variational estimator away from the maximum likelihood estimator.[7] Making the variational posterior covariance more flexible is important to reduce bias, as our results based on simulated data show.

Remark 2. *It is interesting to note that the Gaussian variational family is not restrictive in the linear Gaussian model. Consider as an example the mean-field case where one assumes that*

$$q(z_{1:T} \mid y_{1:T}) = \prod_{t=1}^{T} q_t(z_t \mid y_{1:T}).$$

Suppose one maximizes the ELBO with respect to unrestricted marginal densities $q_1, ..., q_T$, without imposing they are Gaussian. In that case, it can be shown using variational calculus that the q_t's that maximize the ELBO satisfy

$$q_t \propto \exp(\mathbb{E}_{q_{-t}} \log p),$$

where $\mathbb{E}_{q_{-t}} \log p$ is the expectation of the log-posterior density with respect to all variational marginal densities except the one in period t (and $\propto$ denotes "proportional to"). Since the log-posterior is quadratic in $z_1, ..., z_T$ (see (22)), it then follows that q_t is Gaussian. Hence, in a linear Gaussian model, relying on Gaussian variational posteriors is not restrictive, even in a mean-field approach. However, Gaussianity of the variational family may be restrictive in nonlinear models, which we turn to in the next section.

[7] Further, note that while the mean-field approach does not restrict the variational posterior mean, since $\mu_q = \mu_{z|y}$, it forces the variational posterior variances to be equal to the diagonal elements of the precision matrix, $\sigma^2_{q,t} = (\omega_t^{\text{post}})^{-2}$.

5 A Nonlinear Non-Gaussian Model

We now extend the linear Gaussian model in two directions. First, we introduce nonlinearity in the latent process by incorporating a *nonlinear conditional mean* and *state-dependent volatility*. Second, we *relax* the Gaussian assumptions for the initial latent state and the transitory component.

5.1 The Nonlinear Model

We focus on model (1)-(2)-(3), where we allow both the latent mean function $\mu(z_{t-1})$ and the state-dependent volatility function $\sigma(z_{t-1})$ to be nonlinear, while adding non-Gaussian features in both the initial state z_1 and the transitory shock e_t. The nonlinear functions $\mu(z_{t-1})$ and $\sigma(z_{t-1})$ capture deviations from the linear AR(1) process that we studied in the previous section.

To flexibly depart from Gaussianity, we model the transitory shock e_t and the initial latent state z_1 as *sinh–arcsinh* transformations of a standard normal random variable (Jones and Pewsey, 2009). The resulting cumulative distribution functions are

$$\Psi_e(e) = \Phi\big(\sinh\big(\psi_2 \operatorname{asinh}(e/\psi_1)\big)\big)\,, \tag{30}$$

$$F_{z_1}(z) = \Phi\big(\sinh\big(\gamma_2 \operatorname{asinh}(z/\gamma_1)\big)\big)\,, \tag{31}$$

where (ψ_1, ψ_2) and (γ_1, γ_2) are scale and tail-shape parameters for e_t and z_1, respectively, and $\Phi(\cdot)$ denotes the standard normal CDF. This transformation retains zero mean and symmetry but adjusts tail thickness: values $\psi_2 < 1$ (or $\gamma_2 < 1$) generate heavier tails and positive excess kurtosis relative to the Gaussian benchmark, while $\psi_2 > 1$ yields thinner tails. The scale parameters ψ_1 and γ_1 control dispersion, allowing independent tuning of standard deviation and tail behavior. This parameterization provides a tractable way to match symmetric empirical distributions with excess kurtosis.

The resulting model departs from the benchmark state-space setting in several important ways: the presence of nonlinear mean and volatility breaks the linear-Gaussian assumption, the non-Gaussianity in z_1 and e_t implies that the posterior distribution over the latent path $z_{1:T}$ is no longer Gaussian, and its precision matrix is no longer tridiagonal in general. Moreover, the stochastic volatility term $\sigma(z_{t-1})$ introduces multiplicative heteroskedasticity, which creates nonlinearities in the conditional likelihood and undermines the conjugacy that made posterior calculations analytically tractable in the linear Gaussian case.

In this context, our goal is to assess whether a variational inference approach based on a Gaussian variational family – with or without additional structure – is able to correctly recover

Table 2: Parameters for the Simulated DGP in the Nonlinear Model

Parameter	Value	Description
$\mu(z_{t-1})$	$-0.25 + 0.1 \log\left[1 + \exp\left(\frac{1}{0.1}\left(0.9 z_{t-1} + 0.25\right)\right)\right]$	hockey stick shape
$\sigma(z_{t-1})$	$\log\left(1 + \exp\left(-1.55 + 0.35 z_{t-1}^2\right)\right)$	quadratic volatility function
$\psi(e_t)$	$\Phi\big(\sinh\big(0.47\ \mathrm{asinh}(e_t/0.033)\big)\big)$	trans. shock $\sigma_e = 0.16$, $kurt_e = 10.0$
$f(z_1)$	$\Phi\big(\sinh\big(0.89\ \mathrm{asinh}(z_1/0.34)\big)\big)$	initial state $\sigma_{z_1} = 0.40$, $kurt_{z_1} = 3.3$
N	30,000	number of individuals
T	6	number of periods

the model parameters despite the non-Gaussianity of the posterior.

Parameterization. To calibrate the DGP, we choose parameters for the transitory component e_t to match a standard deviation of 0.16 and a kurtosis of 10, and for the initial latent state z_1 to match a standard deviation of 0.4 and a kurtosis of 3.3, in line with Arellano et al. (2017). This yields parameter values $(\psi_1, \psi_2) = (0.033, 0.47)$ for e_t and $(\gamma_1, \gamma_2) = (0.34, 0.89)$ for z_1.

For the nonlinear conditional mean function $\mu(z_{t-1})$, we adopt a specification designed to capture a *hockey stick* shape: the left tail of the distribution is nearly flat (mimicking a floor in earnings changes, e.g., due to minimum wages), while the right side increases with a persistence parameter close to 0.9. Specifically, in the DGP we assume that

$$\mu(z_{t-1}) \;=\; -0.25 + 0.1 \log\left[1 + \exp\left(\tfrac{1}{0.1}\left(0.9 z_{t-1} + 0.25\right)\right)\right], \tag{32}$$

which is strictly increasing and smooth (see Figure 2). We maintain a quadratic innovation volatility function of the form

$$\sigma(z_{t-1}) = \log\left(1 + \exp\left(-1.35 + 0.35 z_{t-1}^2\right)\right). \tag{33}$$

Finally, we simulate $N = 30,000$ individuals over $T = 6$ periods. The parameters of the model are summarized in Table 2.

In estimation, we allow for more flexible functional forms:

$$\mu(z_{t-1}) = \alpha_0 + \alpha_1 \log\left[1 + \exp\left(\tfrac{1}{\alpha_1}\left(\mu_0 + \mu_1 z_{t-1} + \mu_2 z_{t-1}^2 - \alpha_0\right)\right)\right], \tag{34}$$

$$\sigma(z_{t-1}) = \log\left[1 + \exp\left(\sigma_0 + \sigma_1 z_{t-1} + \sigma_2 z_{t-1}^2\right)\right], \tag{35}$$

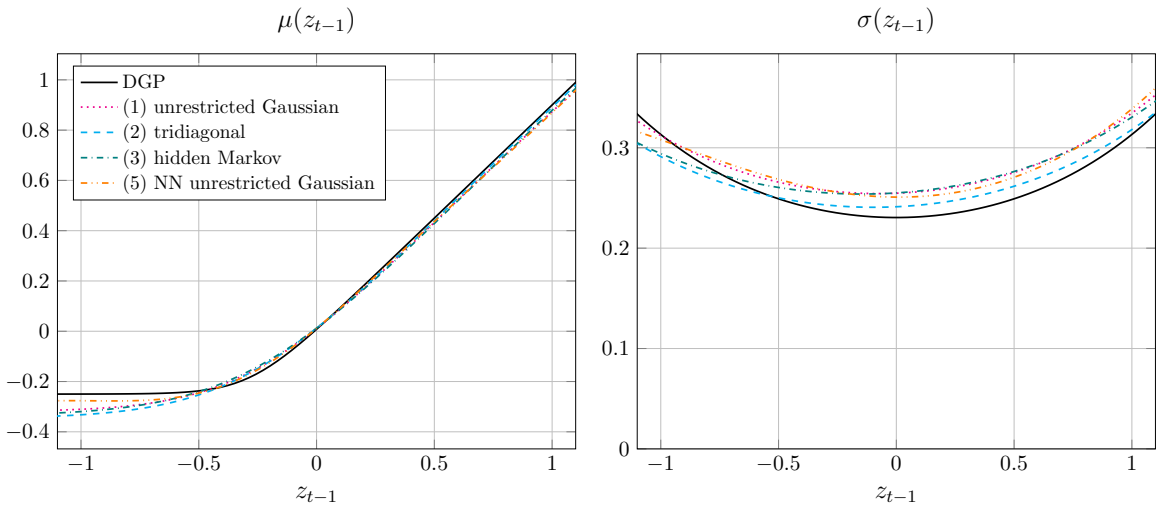

Figure 2: Simulation Results in the Nonlinear Model

Note: The figure plots the conditional mean and volatility functions in equation (2) implied by the simulated nonlinear DGP, together with their estimated counterparts obtained under four variational posterior specifications. Each panel shows the true conditional mean or volatility (solid lines), as specified in (32)–(33), and their estimated counterparts (dashed lines) across the support of z_{t-1}.

and compare the performance of the same set of variational posteriors as in the benchmark model. In all cases, we optimize the evidence lower bound (ELBO) using reparameterization-based gradient ascent.

5.2 Results for the Nonlinear Model

Figure 2 and Table 3 summarize the estimation results for the nonlinear specification. The unrestricted Gaussian posterior, as well as the structured Markov posterior, capture both the central tendency and dispersion of the latent persistent component z_t. The tridiagonal approximation, while slightly less flexible, performs similarly. All three approximations recover the true parameters of the process of z_t quite well, including the curvature in the conditional volatility function, as can be seen from the first three rows of Table 3.

All these variational estimators are based on model (34)-(35) that has a parsimonious parametric structure. In applications of variational inference in machine learning, it is common to entertain much richer specifications based on neural networks. In the next-to-last row of Table 3, and in Figure 2, we report estimates based on such a specification where $\mu(\cdot)$ and $\log \sigma(\cdot)$ are specified as feedforward neural networks. We can see that, for the sample size that we consider ($N = 30,000$ and $T = 6$), this highly flexible model gives similar estimates to the parametric specification (34)-(35).

Table 3: Simulation Results in the Nonlinear Model

Parameter	α_0	α_1	μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	σ_{z_1}	kurt_{z_1}	σ_e	kurt_e
DGP	-0.25	0.10	0.00	0.90	0.00	-1.35	0.00	0.35	0.40	3.3	0.16	10.0
Variational posterior												
(1) unrestricted Gaussian	-0.30	0.23	-0.06	1.00	-0.06	-1.26	0.03	0.24	0.41	3.3	0.14	3.0
(2) tridiagonal	-0.30	0.19	-0.03	0.94	-0.03	-1.28	0.06	0.23	0.41	3.3	0.14	2.9
(3) hidden Markov	-0.33	0.33	-0.14	1.11	-0.09	-1.24	0.05	0.19	0.41	3.3	0.14	3.1
(4) diagonal	-0.26	0.19	-0.04	0.94	-0.06	-1.12	0.07	0.23	0.42	3.3	0.08	3.0
Neural Network model												
(5) unrestricted Gaussian	-0.27	0.14	-0.01	0.89	0.00	-1.26	0.08	0.22	0.41	3.5	0.14	3.1
Ignoring transitory shocks												
(6)	-0.24	0.19	-0.04	0.92	-0.08	-1.07	0.09	0.22	0.43	3.4	–	–

Note: The table reports the parameter values used in the simulated nonlinear DGP and the corresponding estimates obtained under five variational posterior specifications. The estimation model follows equations (34)–(35), which allow for nonlinear conditional mean and state-dependent volatility functions, and is applied to data generated with sinh–arcsinh innovations for z_1 and e_t.

In contrast, when using a mean-field variational posterior with a diagonal Gaussian structure (fourth row in the table) or using a model that ignores transitory shocks (last row), we find that the approximation fails to recover the mean and volatility functions accurately. In both cases, the estimated slope of the conditional mean function is attenuated relative to the true process, resulting in a downward bias in persistence.

The same two specifications also perform poorly in recovering the conditional variance of the latent process. Both produce variance estimates that are too high across the state space relative to the true state-dependent volatility. By comparison, the other variational approximations we consider, including the unrestricted Gaussian, the tridiagonal precision matrix specification, and the conditional Markov specification, match the true conditional volatility function quite well over the range of latent states, albeit not perfectly.

These findings highlight an important practical insight: even when the true posterior is not Gaussian, variational approximations based on a Gaussian family can remain effective in recovering the parameters of the process of interest. At the same time, restrictive independence assumptions in the posterior or failure to incorporate transitory components can lead to substantial distortions in both persistence and volatility.

However, performance is not uniformly good across parameters. Focusing on the last two columns of Table 3, we see that the unrestricted and structured Gaussian variational specifications slightly underestimate the variance of transitory shocks, and more importantly that they

fail at capturing the high transitory kurtosis. In fact, all methods lead to the same – incorrect – conclusion that transitory shocks are approximately normal. This highlights another important insight regarding variational inference in this setting, as Gaussian specifications lead to incorrect estimation of features of the distribution of transitory shocks. In Section 8 we will explore several approaches to alleviate this issue.

Lastly, regarding computational cost, the nonlinear model is estimated in 3 minutes and 30 seconds.[8] Note that this computational cost has to be assessed in view of the sample size of $30,0000$ individual units and 6 periods.

5.3 Posterior Distributions

To better understand the behavior of different variational posterior families, we examine the posterior distribution $p(z_{1:T} \,|\, y_{1:T})$ directly in a simple two-period setting with $T = 2$. For this exercise, we set the model parameters to their true values, and we fix observations at $(y_1, y_2) = (-0.1, 0.1)$. We then compare the exact posterior $p(z_1, z_2 \,|\, y_1, y_2)$ with the best fit within each variational family, obtained by minimizing the KL divergence between the true posterior and the approximating distribution. In addition to showing the overall posterior density, the contour plots in Figure 3 also report the mean of the distribution together with the associated covariance ellipse whose axes pass through the posterior mean (the black dot in the plot). The orientation of the major axis reflects the correlation between z_1 and z_2: a tilted ellipse indicates that the posterior recognizes the dependence across periods, while a vertical or horizontal axis would imply independence.

In the top panel of Figure 3 we show posterior contour plots in the benchmark linear Gaussian model. Both the unrestricted Gaussian and hidden Markov specifications capture the posterior shape well. Both variational posteriors also reproduce the tilt of the true ellipse, thereby capturing the covariance structure accurately. By contrast, the diagonal normal posterior effectively forces the correlation to zero. This illustrates how mean-field restrictions systematically understate the dependence in the latent dynamics, consistently with what we documented in Section 4.

In the bottom panel of Figure 3 we show posterior contour plots for the nonlinear model, whose true posterior takes on a distinct stretched kite shape. This geometry reflects the state-dependent volatility, which stretches the posterior mass along both the vertical and horizontal directions. The unrestricted Gaussian variational posterior cannot reproduce this structure,

[8] This is using an NVIDIA H100 graphic unit with 80GB of RAM. Memory usage is 2.5%, and CPU use is 3%.

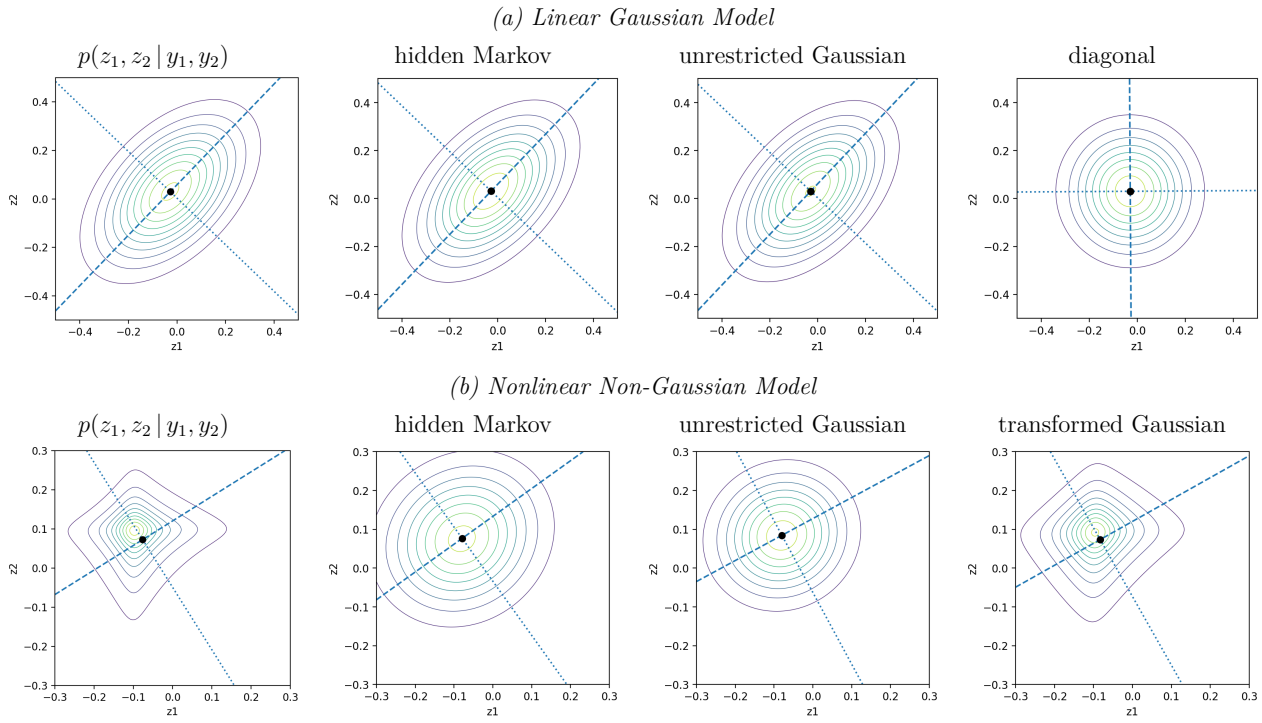

Figure 3: Variational Approximation to the True Posterior Density $p(z_{1:T} \mid y_{1:T})$

Note: The figure displays contour plots of the true posterior density $p(z_1, z_2 \mid y_1, y_2)$ and its variational approximations for a two-period setting with $(y_1, y_2) = (-0.1, 0.1)$. The top panel corresponds to the linear Gaussian model, and the bottom panel to the nonlinear non-Gaussian model. Each column shows a different variational posterior specification. In every panel, the black dot marks the posterior mean and the ellipse illustrates the corresponding covariance matrix, whose principal axes pass through the mean.

approximating it instead with tilted ellipses, while the hidden Markov variational posterior also remains misaligned with the true contours. Looking at the mean and covariance, the true posterior has its mode shifted to the top-left of the mean, indicating skewness that is not captured by either the hidden Markov or the unrestricted Gaussian approximations. While these two families succeed in reproducing the overall covariance ellipse correctly, they miss the asymmetry in the placement of the mode.

This example underscores the limits of Gaussian variational approximation. When the true posterior departs from elliptical shapes, neither the hidden Markov structure nor the unrestricted Gaussian specification are sufficient, and richer approximating families that allow for nonlinear distortions of the latent space are needed to capture the geometry. In the bottom right graph of Figure 3 we report contour plots of a transformed Gaussian variational posterior (based on normalizing flows), which we will describe formally in Section 8. This approximation

performs markedly better by matching the kite-shaped geometry of the true posterior and capturing both skewness and dependence. This combination of accurate geometry and alignment of higher-order moments highlights the appeal of richer posterior families in nonlinear settings.

6 Extensions: Serial Correlation and Heterogeneity

In this section, we show how the variational approach can be easily modified to handle two important extensions of the baseline nonlinear model incorporating serially correlated transitory shocks and time-invariant heterogeneity.

6.1 Serially Correlated Transitory Shocks

We first extend the earnings dynamics model to allow for serial correlation in the transitory shock e_t. Allowing for an MA(1) component in the transitory shock directly relates to the empirical strategy of Meghir and Pistaferri (2004), who emphasize the importance of serial correlation in transitory shocks for accurately characterizing the dynamics of income volatility. Specifically, we introduce an MA(1) structure in the transitory component e_t, while maintaining the Markovian evolution of the latent persistent component z_t. The data-generating process is specified as:

$$y_t = z_t + e_t, \tag{36}$$

$$z_t = \mu(z_{t-1}) + \sigma(z_{t-1})u_t, \tag{37}$$

$$e_t = \varepsilon_t + \zeta\varepsilon_{t-1}, \tag{38}$$

$$z_1 \sim f_\alpha, \quad u_t \sim \mathcal{N}(0,1), \quad \varepsilon_t \sim \psi_\gamma. \tag{39}$$

The new key feature is the presence of serial dependence in the transitory shock process: each observed income y_t now depends not only on the current latent state z_t, but also indirectly on the past component ε_{t-1}. This implies that y_t is serially correlated even after conditioning on the latent state z_t. As a result, the likelihood function is no longer conditionally independent across time given the latent states, and instead exhibits overlapping dependencies across multiple periods. The likelihood function in this model still involves integrating over the latent process $z_{1:T}$, see (4), however the form of the conditional density $\psi_\gamma(y_{1:T} - z_{1:T})$ is now more complex due to the presence of the autocorrelated transitory component.

In models with serially independent transitory shocks, such as those considered in the previous sections, the structured variational posterior of the form (28) provides a faithful approximation to the true posterior. This structure is motivated by the fact that, in a first-order Markov

latent process with conditionally independent observations, the distribution of z_t given the past latent state z_{t-1} and the future observations $y_{t:T}$ captures all relevant dependencies. However, when the transitory component follows an MA(1) process, the observation y_t depends not only on the current latent state z_t, but also on the lagged shock ε_{t-1}. This term is only partially revealed through y_{t-1}, which itself contains ε_{t-2}. As a result, the density $p(z_t \mid z_{t-1}, y_{1:T})$ generally depends on a backward window of past observations $y_{t-L:t-1}$ with $L \geqslant 1$, together with future observations $y_{t:T}$. The information in these past observations cannot be fully mediated through z_{t-1} alone, and omitting them leads to a posterior that understates the dependence between the latent state and the serially correlated noise.

To reflect this structure, we propose to specify a variational posterior that conditions each latent state on its lag z_{t-1}, and on a low-dimensional residual r_t that captures relevant past information:

$$q(z_{1:T} \mid y_{1:T}) = q(z_1 \mid y_{1:T}) \prod_{t=2}^{T} q(z_t \mid z_{t-1}, y_{t:T}, r_t), \tag{40}$$

where the residual is defined recursively as

$$r_t = y_t - z_t - \zeta\, r_{t-1}, \quad r_1 = y_1 - z_1. \tag{41}$$

Including r_t in the conditioning set enables the variational posterior to adapt to the dependence induced by the MA(1) structure and any non-Gaussianity in ε_t, while keeping the approximation computationally tractable.

We report simulation-based evidence on this approach and other variational specifications for a model with serially-correlated transitory shocks in Appendix A.1. We find that the unrestricted Gaussian and the structured approach based on (40)-(41) recover the latent state process quite well. However, as in the case with independent shocks, the variational approach is not able to capture the kurtosis of transitory shocks.

6.2 Time-Invariant Heterogeneity

We next augment the nonlinear model to allow for a time-invariant latent type a. We postulate the following model

$$y_t = z_t + e_t, \tag{42}$$

$$z_t = \mu(z_{t-1}, a) + \sigma(z_{t-1}, a) u_t, \tag{43}$$

$$(a, z_1) \sim f_\alpha, \quad u_t \sim \mathcal{N}(0,1), \quad e_t \mid a \sim \psi_\gamma(\cdot; a), \tag{44}$$

where u_t are serially independent, independent of e_t, and independent of a. The type a can be vector-valued, and follows a joint distribution f_α together with the initial condition z_1.

Let $\theta = (\alpha, \mu(\cdot,\cdot), \sigma(\cdot,\cdot))$, and let f_θ denote the joint density of $a, z_1, ..., z_T$. The log-likelihood function is

$$\mathcal{L}_{\theta,\gamma}(y_{1:T}) = \log \int f_\theta(a, z_{1:T})\, \psi_\gamma(y_{1:T} - z_{1:T}; a)\, dz_{1:T}\, da, \tag{45}$$

where now the integral is taken with respect to $(a, z_1, ..., z_T)$. Given a variational posterior density $q_\phi(a, z_{1:T} \mid y_{1:T})$, the ELBO is given by

$$\mathcal{E}_{\theta,\gamma,\phi}(y_{1:T}) = \mathbb{E}_{q_\phi(a,z_{1:T}\mid y_{1:T})}\left[\log \frac{f_\theta(a, z_{1:T})\psi_\gamma(y_{1:T} - z_{1:T}; a)}{q_\phi(a, z_{1:T} \mid y_{1:T})}\right]. \tag{46}$$

We report simulations for a nonlinear model with latent heterogeneity in Appendix A.2. In the model, the variance of transitory shocks is heterogeneous across individuals, as in Almuzara (2020). We find that the unrestricted Gaussian approach achieves a good approximation to the mean and, to a lesser extent, to the volatility of the latent state process. However, the variational approach is again not able to capture features of transitory shocks such as kurtosis, and it does not accurately capture the individual heterogeneity in transitory variances.

7 Nonlinear Earnings Processes in the PSID

We now apply variational inference to real-world earnings data. Our objective is to compare the earnings process estimated using variational inference to the one reported in Arellano et al. (2017), who study earnings processes using a nonlinear latent variable model with a quantile-based stochastic EM approach. Relative to this paper, we rely on a specification for annual income (as opposed to biennial) with serially correlated transitory shocks and estimate the model using variational inference either.

7.1 Estimated Earnings Process

We estimate the nonlinear model introduced in Section 2 using data from the Panel Study of Income Dynamics (PSID) covering the period 1980–1990. The PSID is a long-running U.S. household survey with detailed annual information on labor earnings, hours worked, employment status, and demographic characteristics. Following Blundell et al. (2008), we restrict the sample to male household heads aged 25–60 with strong labor force attachment, excluding the self-employed, individuals reporting fewer than 520 annual hours, and those with missing or

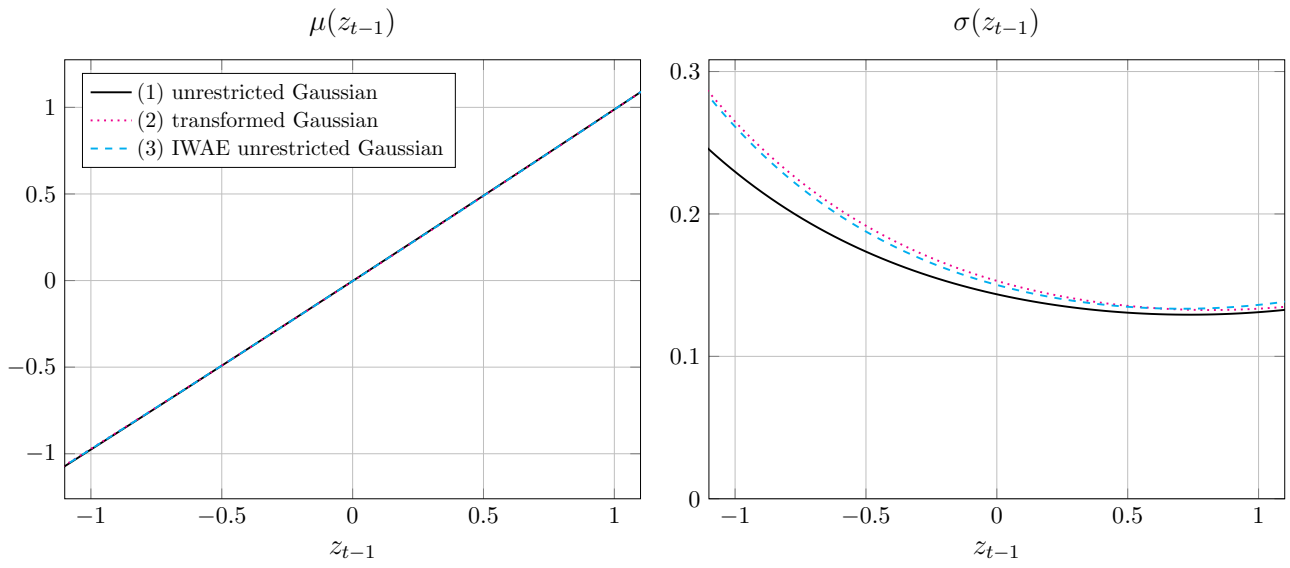

Figure 4: Estimates on the PSID

Note: The figure plots the estimated conditional mean $\mu(z_{t-1})$ and conditional volatility $\sigma(z_{t-1})$ of the latent earnings component z_t using PSID data from 1980–1990. Solid lines correspond to estimates obtained with an unrestricted Gaussian variational posterior. Dashed lines show estimates based on alternative variational families described in Section 8.

implausible income reports. Labor income is constructed net of transfers and taxes using the PSID family files.[9]

Our specification allows for nonlinear conditional mean $\mu(z_{t-1})$ and volatility $\sigma(z_{t-1})$ in the latent process, as well as flexible initial condition z_1 and transitory shock e_t distributions, including a moving average specification for transitory shocks. Both $\mu(z_{t-1})$ and $\sigma(z_{t-1})$ are specified using quadratic polynomials as in (34)-(35). We incorporate an MA(1) component in the transitory innovation to capture serial correlation in transitory shocks.

Our main estimates are based on an unrestricted Gaussian variational posterior family. The results can be found in Figure 4 (in solid lines) and Table 4 (in the first row). Our estimates reveal a highly persistent latent earnings process. The conditional mean $\mu(z_{t-1})$ is close to linear, with an autoregressive coefficient of 0.98. The conditional volatility $\sigma(z_{t-1})$ exhibits a tilted U-shape: volatility is high at the lower end of the distribution, decreases toward the middle, and rises again – though more modestly – at the upper end.

These estimates imply that persistence in z_t is nonlinear. As in Arellano et al. (2017), we measure persistence as the derivative of the conditional quantile function of z_t given z_{t-1} with

[9]We apply the procedure from the replication package in Blundell et al. (2008), where log household earnings are residualized on a set of demographic variables in a first step.

respect to z_{t-1}. In words, persistence measures how the current earnings component z_t changes when the past component z_{t-1} changes, for given values of the latter and the shock u_t. In our conditionally Gaussian model the quantile function is

$$Q_\tau(z_t \,|\, z_{t-1}) = \mu(z_{t-1}) + \sigma(z_{t-1})\Phi^{-1}(\tau), \quad \text{for all } \tau \in (0,1),$$

where Φ is the standard normal cdf. We then measure persistence as

$$\begin{aligned} \rho(z_{t-1}, \tau) &= \nabla_{z_{t-1}} Q_\tau(z_t \,|\, z_{t-1}) \\ &= \underbrace{\nabla_{z_{t-1}} \mu(z_{t-1})}_{\text{state-dependent mean}} + \underbrace{\nabla_{z_{t-1}} \sigma(z_{t-1})\Phi^{-1}(\tau)}_{\text{state-dependent volatility}}. \end{aligned} \tag{47}$$

According to our estimates, the state-dependent mean component is close to a constant as $\mu(z_{t-1})$ is approximately linear. In contrast, the state-dependent volatility component is U-shaped in z_{t-1}, which implies that the persistence measure $\rho(z_{t-1}, \tau)$ depends both on the state z_{t-1} and the shock τ (i.e., the percentile rank of u_t).

We report our estimate of the persistence surface in Panel (a) of Figure 5. The two horizontal axes indicate the percentile rank of z_{t-1} and the percentile rank τ of u_t, respectively, and the vertical axis indicates the values of $\rho(z_{t-1}, \tau)$. We see that persistence is approximately constant, and close to $\rho = 1$, for central values of z_{t-1} and u_t. However, high-u_t shocks for low-z_{t-1} households are associated with a lower persistence, as low as 0.5. We also observe that low-u_t shocks for high-z_{t-1} households are also associated with a lower ρ, although the decrease is not as stark. Lastly, our estimates indicate some increase in ρ for high-u_t/high-z_{t-1}, and especially low-u_t/low-z_{t-1} combinations. On the graph on the right of the figure, we reproduce the persistence estimates from Arellano et al. (2017). While the two sets of estimates were obtained from different specifications and estimation methods, and using different samples[10], they tend to agree to a large extent.

In the last five columns of Table 4 we report our estimates of the other parameters of the model, corresponding to the initial state z_1 and the transitory shocks e_t. In particular, we find that transitory shocks are positively correlated, with a moving average coefficient equal to 0.22. Our estimates also indicate that transitory shocks have excess kurtosis relative to the normal, although based on the simulation evidence in the previous sections it is likely that the kurtosis of e_t might be understated by our Gaussian variational estimator.

As a way to probe the robustness of the results, we also report estimates based on two alternative approaches: one that uses a transformation of the Gaussian as a variational family,

[10] Arellano et al., 2017 use biennial post-1999 PSID data.

Table 4: Estimates on the PSID

Parameter	μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	σ_{z_1}	kurt_{z_1}	σ_ϵ	kurt_ϵ	ζ
Variational posterior											
(1) unrestricted Gaussian	0.00	0.98	0.01	-1.88	-0.31	0.21	0.53	4.2	0.18	5.0	0.22
(2) transformed Gaussian	0.00	0.98	0.01	-1.81	-0.38	0.23	0.53	4.2	0.16	4.3	0.15
(3) IWAE unrestricted Gaussian	0.00	0.98	0.01	-1.83	-0.36	0.25	0.53	4.1	0.17	4.7	0.20

Note: The table reports parameter estimates for the nonlinear earnings model estimated on the PSID from 1980–1990. The first row corresponds to the unrestricted Gaussian variational posterior, while subsequent rows report estimates obtained using different approaches described in Section 8. Reported parameters include those of the conditional mean $\mu(z_{t-1})$, conditional volatility $\sigma(z_{t-1})$, the initial state z_1, and the transitory component e_t, including its moving-average coefficient ζ.

and an approach based on using the variational density as a proposal for importance sampling. We will provide details about these two approaches in Section 8. The estimates shown in Figure 4 (in dashed lines) and Table 4 (in the second and third rows) are similar to the ones based on our baseline Gaussian variational approach. One difference is that the two alternative methods imply slightly more curvature in the conditional volatility $\sigma(z_{t-1})$.

7.2 Certainty Equivalent and Risk Premium

To assess the economic relevance of the features of the earnings process that we have estimated, we next report several summaries that quantify (under certain assumptions) the risk faced by households. To proceed, let $z_t \sim f_t(z_t \mid z_1)$ denote the distribution of future outcomes conditional on the initial latent state z_1, and let $\beta \in (0, 1)$ denote the discount factor. Let $u(z_t) = U(e^{z_t})$ denote household utility. Here we interpret the persistent earnings component z_t as a proxy for log-consumption.[11]

The household's expected utility is given by

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t u(z_t)\right] = \sum_{t=0}^{\infty} \beta^t \int u(z) f_t(z \mid z_1)\, dz. \tag{48}$$

Rewriting this expression by exchanging the order of summation and integration yields:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t u(z_t)\right] = \int u(z) \left(\sum_{t=0}^{\infty} \beta^t f_t(z \mid z_1)\right) dz. \tag{49}$$

[11] Note that the utility function does not vary over time.

(a) Estimates using Variational Inference *(b) Estimates from Arellano et al. (2017)*

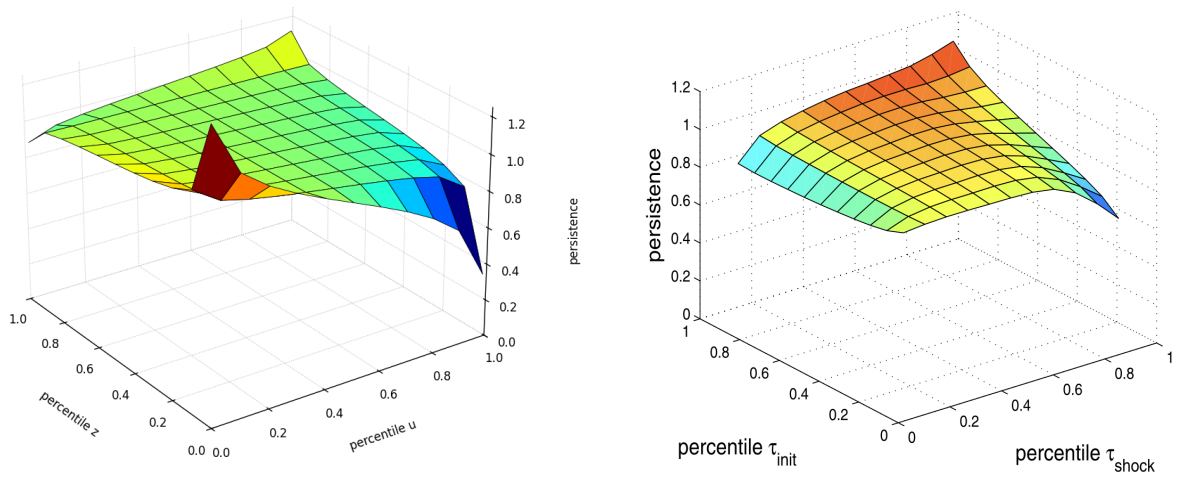

Figure 5: Estimates of Nonlinear Persistence of z_t on the PSID

Note: The figure displays estimates of the state- and shock-dependent persistence measure $\rho(z_{t-1}, \tau)$ defined in equation (47). Panel (a) shows the persistence surface estimated using variational inference, while Panel (b) reproduces the corresponding estimates from Arellano et al. (2017). The horizontal axes indicate the percentile ranks of z_{t-1} and of the innovation u_t, and the vertical axis reports the implied persistence $\rho(z_{t-1}, \tau)$.

We define the *discounted mixture density* as

$$\widetilde{f}(z \mid z_1) = (1-\beta) \sum_{t=0}^{\infty} \beta^t f_t(z \mid z_1), \tag{50}$$

which integrates to one and thus defines a valid probability distribution.

Given $\widetilde{f}(y \mid z_1)$, and for any continuous, integrable utility function $U(\cdot)$, expected discounted sums of utility can be computed in closed form as a single integral:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t u(z_t)\right] = \frac{1}{1-\beta} \int u(z) \widetilde{f}(z \mid z_1)\, dz = \frac{1}{1-\beta} \mathbb{E}_{\widetilde{f}}[u(z)]. \tag{51}$$

For example, the *certainty equivalent* c^{CE} defined as the solution to

$$\sum_{t=0}^{\infty} \beta^t u(c^{CE}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t u(z_t)\right], \tag{52}$$

can be obtained as

$$c^{CE} = u^{-1}\left(\mathbb{E}_{\widetilde{f}}[u(z)]\right). \tag{53}$$

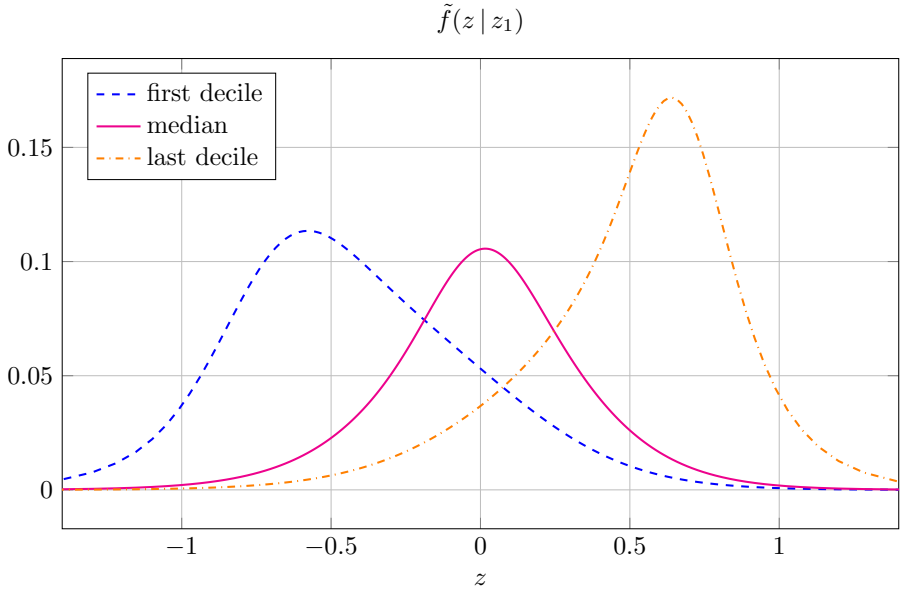

Figure 6: Discounted Mixture Density $\widetilde{f}$ in the PSID

Note: The figure plots the discounted mixture density $\widetilde{f}(z \mid z_1)$ defined in equation (50) for $\beta = 0.9$, estimated from the PSID over 1980–1990. The horizontal axis measures the latent earnings component z, and the vertical axis reports the corresponding discounted probability density conditional on the initial state z_1.

In turn, the *risk premium* – defined as the amount an agent would be willing to pay to avoid uncertainty – can be obtained as follows:

$$\pi = 1 - \frac{c^{CE}}{\mathbb{E}_{\widetilde{f}}[e^z]}. \tag{54}$$

The discounted density $\widetilde{f}$ estimated on the PSID data (Figure 6) inherits the asymmetric features of the nonlinear model of Section 5. The conditional distribution of discounted log earnings is skewed: households starting from the lower end of the income distribution face right-skewed earnings innovations, whereas those starting from the upper end experience left-skewed distributions. In addition, the peak of $\widetilde{f}$ is more pronounced at the top than at the bottom of the distribution, reflecting the fact that the conditional volatility increases more sharply at low income levels, while remaining comparatively muted at higher levels of the distribution.

In Figure 7 we report certainty equivalents and risk premia estimated from the PSID for $\beta = 0.9$ under quadratic, logarithmic, and CRRA utility (with parameter 2.0). The certainty equivalent is increasing and convex in the initial latent state, reaching values at the top of the distribution that are roughly five times higher than those at the bottom. In addition, we find

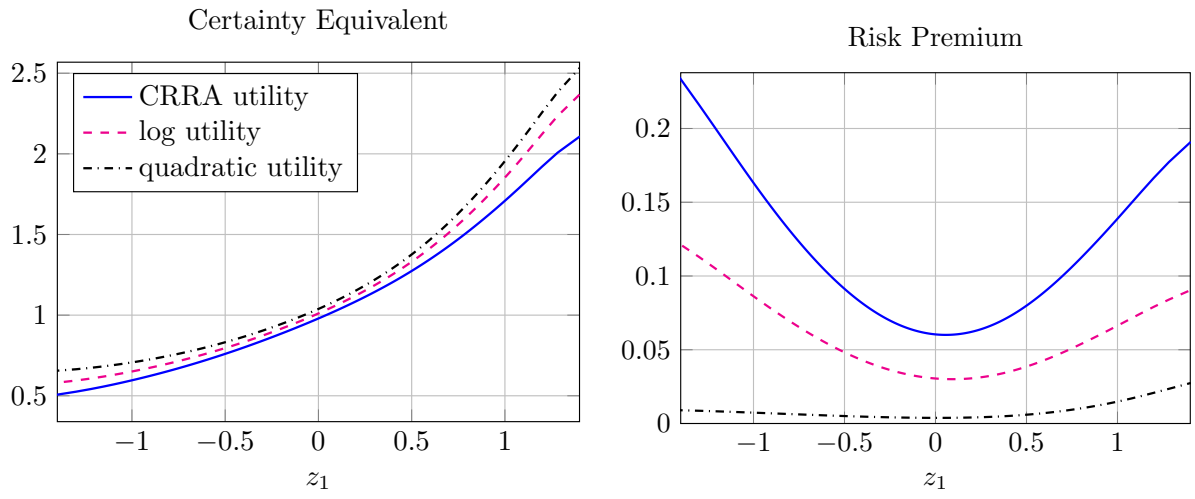

Figure 7: Certainty Equivalent and Risk Premium in the PSID

Note: The figure displays certainty equivalents (left) and risk premia (right) computed from PSID estimates of the nonlinear earnings model for $\beta = 0.9$, using three alternative utility specifications: quadratic, logarithmic, and CRRA. Each panel reports values as a function of the initial latent state z_1, based on the estimated discounted mixture density $\widetilde{f}(z \mid z_1)$.

that risk premia display substantial heterogeneity across the distribution. For example, under log utility they range from about 2 percent in the middle of the distribution to 12 percent at the lower end. With more curvature in preferences, as under CRRA utility, risk premia become even larger, exceeding 20 percent for households starting at the lower tail of the distribution.

8 Two Alternatives to Gaussian Variational Inference

Our results based on simulated data indicate that, while Gaussian variational approximations tend to recover the conditional mean and variance quite well, the method fails at capturing the kurtosis of transitory shocks. A number of strategies have been proposed in the literature to improve the accuracy of the approximation. Here we briefly review two of these strategies.

8.1 Normalizing Flows

Originally proposed by Rezende and Mohamed (2015), a normalizing flow is an invertible transformation h that maps a simple base distribution (e.g. a standard Gaussian) into a complex distribution. By incorporating flows into variational inference, we can start with a Gaussian $q_0(z)$ as base density, and define a sequence of transformations $z_K = h_K \circ \cdots \circ h_1(z_0)$. The resulting density $q_K(z_K)$ is computed using the change-of-variables formula. By choosing trans-

formations h_k with tractable Jacobians and expressive functional forms, the variational family can capture complex density shapes, including skewness, heavy tails, and multimodality. One practical advantage is that flows maintain computational tractability (the transformations are chosen so that Jacobian determinants are easy to compute), so the ELBO with a flow-based $q_K(z_K)$ can still be optimized efficiently.

To capture non-Gaussian features of the posterior distribution such as skewness and excess kurtosis, we consider a one-dimensional transformation of a standard normal variable based on the sinh–arcsinh family introduced in Section 5. This family allows us to preserve the tractability and reparameterization benefits of the Gaussian while introducing controlled deviations from normality. Let $\widetilde{z} \sim \mathcal{N}(0,1)$ be a standard normal random variable. We define the transformed latent variable z as

$$z = T_\lambda(\widetilde{z}) = \mu + \sigma \cdot \sinh\left(\frac{\operatorname{arsinh}(\widetilde{z}) + \epsilon}{\delta}\right), \tag{55}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ control the location and scale, $\delta > 0$ governs the kurtosis, $\epsilon \in \mathbb{R}$ introduces skewness, and $\lambda = (\mu, \sigma, \epsilon, \delta)$. When $\delta = 1$ and $\epsilon = 0$, the transformation reduces to the identity and $z \sim \mathcal{N}(\mu, \sigma^2)$. The reparameterization trick applies directly, as z is obtained by a differentiable transformation of a base normal variable. Gradients of the ELBO with respect to λ can then be estimated via differentiation. All terms are available in closed form, and the gradients are computed efficiently.

In Figure 3 we have seen that the transformed Gaussian specification captures the non-Gaussian shape of the posterior density well. At the same time, the sinh–arcsinh family has a specific, potentially restrictive functional form. The transformation is applied element-wise and does not model correlations across latent dimensions. Parsimonious transformations of the Gaussian appear promising to improve the estimation of earnings dynamics models given the presence of nonlinearity and non-Gaussian features in earnings data.

8.2 Importance-Weighted Variational Inference

Another approach is to use the fitted Gaussian variational posterior as a proposal density for *importance sampling*. This strategy is referred to as Importance-Weighted Autoencoders (IWAE) (Wu et al., 2016, Cremer et al., 2017, Kim and Mnih, 2020). It exploits multiple draws from the variational posterior to tighten the ELBO, thereby reducing the gap with the true log-likelihood.

Formally, let $q_\phi(z_{1:T} \mid y_{1:T})$ denote a variational posterior family with parameter ϕ, and suppose we draw K independent values $\{z^{(k)}\}_{k=1}^K \sim q_\phi(z \mid y)$. The importance-weighted ELBO

is defined as

$$\mathcal{E}^{(K)}_{\theta,\gamma,\phi}(y_{1:T}) = \mathbb{E}_{z^{(1)},\ldots,z^{(K)}\sim q_\phi(z_{1:T}\,|\,y_{1:T})}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{f_\theta(z^{(k)}_{1:T})\psi_\gamma(y_{1:T}-z^{(k)}_{1:T})}{q_\phi(z^{(k)}_{1:T}\,|\,y_{1:T})}\right)\right],$$

where $K \geqslant 1$ controls the tightness of the bound. For $K = 1$, this recovers the standard ELBO; larger K values yield tighter bounds that approach the true log-likelihood from below. Indeed, it follows from Jensen's inequality that

$$\underbrace{\mathcal{E}_{\theta,\gamma,\phi}(y_{1:T})}_{\text{ELBO}} = \mathcal{E}^{(1)}_{\theta,\gamma,\phi}(y_{1:T}) \leqslant \ldots \leqslant \mathcal{E}^{(K)}_{\theta,\gamma,\phi}(y_{1:T}) \leqslant \mathcal{E}^{(K+1)}_{\theta,\gamma,\phi}(y_{1:T}) \leqslant \ldots \leqslant \underbrace{\mathcal{L}_{\theta,\gamma}(y_{1:T})}_{\text{log-likelihood}}.$$

Diagnostics. The normalized importance weights also allow the researcher to compute two diagnostics at essentially no additional cost. The *effective sample size* (ESS) measures how many independent draws from the true posterior are effectively represented by the draws used in importance sampling. Formally, given the normalized importance weights

$$\widetilde{w}_k = \frac{w_k}{\sum_{j=1}^{K} w_j}, \qquad \text{with} \qquad w_k = \frac{f_\theta(z^{(k)}_{1:T})\psi_\gamma(y_{1:T}-z^{(k)}_{1:T})}{q_\phi(z^{(k)}_{1:T}\,|\,y_{1:T})},$$

where $z^{(k)}$ are independent draws from the variational posterior q_ϕ, the ESS is defined as

$$\text{ESS} = \frac{1}{K\sum_{k=1}^{K}\widetilde{w}_k^2}.$$

This quantity lies between zero and one. An ESS close to one indicates that the variational posterior aligns well with the true posterior, since all samples contribute evenly to the importance-weighted estimate. Conversely, an ESS close to zero indicates severe degeneracy, with nearly all weights concentrated on a single period, implying that the approximation q fails to cover the posterior adequately.

Another useful diagnostic for assessing the quality of the variational approximation is the *ELBO gap* between the evidence lower bound and the log-likelihood,

$$\mathcal{G}_{\theta,\gamma,\phi}(y_{1:T}) = \mathcal{L}_{\theta,\gamma}(y_{1:T}) - \mathcal{E}_{\theta,\gamma,\phi}(y_{1:T}).$$

Computing the ELBO gap $\mathcal{G}_{\theta,\gamma,\phi}(y_{1:T})$ exactly is often infeasible, since it requires evaluating the log-likelihood $\mathcal{L}_{\theta,\gamma}(y_{1:T})$. However, unbiased Monte Carlo estimators of $\mathcal{L}_{\theta,\gamma}(y_{1:T})$ can be constructed using importance sampling, using the same draws as in the IWAE approach. These estimators are considerably more expensive to compute than the standard ELBO. As a result, the ELBO gap is rarely used as an objective for optimization. Instead, it is most useful as an *ex-post* diagnostic.

9 Conclusion

While a growing body of evidence highlights the relevance of nonlinear features in earnings dynamics, nonlinear state-space models remain challenging to estimate. In this paper we explore whether variational inference can provide a reliable alternative to existing methods. We propose a flexible framework that nests the canonical linear Gaussian process while accommodating nonlinear persistence, state-dependent volatility, serially correlated or heavy-tailed transitory shocks, and latent time-invariant heterogeneity. We rely on variational approximations to the posterior distribution of latent states, based on Gaussian specifications, to estimate various versions of the model.

Our simulation results show that Gaussian variational inference recovers the main features of the earnings process quite well, with some important exceptions. The conditional mean and volatility of the persistent component are well estimated across a range of specifications, although some biases are apparent in the various nonlinear models we estimate. Importantly, higher-order moments of transitory shocks such as excess kurtosis remain difficult to capture. Among posterior families, mean-field approximations perform poorly. Applying the method to PSID data, we find a nearly linear conditional mean (with average persistence close to unity), a tilted U-shaped conditional variance across the income distribution, and evidence of serial correlation in transitory shocks. As in Arellano et al. (2017), persistence is lower for high-earnings households experiencing negative shocks and low-earnings households experiencing positive shocks.

Taken together, these findings motivate further study of variational inference as a tractable alternative to traditional likelihood-based methods to estimate realistic models of earnings dynamics. The approach scales well to long panels, is compatible with modern optimization frameworks, and retains flexibility to model key nonlinearities that matter empirically. At the same time, our analysis highlights ongoing challenges, particularly in capturing the shape of the transitory shock distribution.

Future research could extend this agenda in several directions. An important limitation is the lack of theoretical guarantees. Although variational inference has been theoretically justified in some settings, the biases we uncover using simulated data suggest that a Gaussian family does not lead to consistent estimators in the nonlinear models we study. This motivates further work on extensions of the Gaussian approach, such as based on normalizing flows, to study their performance both in practice and in theory.

A key advantage of variational inference is its computational tractability. It provides a

unified, gradient-based estimation framework that can be implemented in modern automatic differentiation environments. It scales efficiently in both the cross-sectional and temporal dimensions of the data, and it remains stable even when the number of latent variables is large. These features make it feasible to estimate flexible nonlinear models not only in survey data such as the PSID but also in large-scale administrative datasets with richer earnings histories. It is our hope that this approach can help uncover new empirical patterns of dynamics and heterogeneity in earnings data.

References

Abowd, J. M. and Card, D. (1989). On the covariance structure of earnings and hours changes. *Econometrica*, 57(2):411–445.

Almuzara, M. (2020). Heterogeneity in transitory income risk. Technical report, Working paper.

Arellano, M. (2014). Uncertainty, persistence, and heterogeneity: A panel data perspective. *Journal of the European Economic Association*, 12(5):1127–1153.

Arellano, M., Blundell, R., and Bonhomme, S. (2017). Earnings and consumption dynamics: A nonlinear panel data framework. *Econometrica*, 85(3):693–734.

Arellano, M., Blundell, R., Bonhomme, S., and Light, J. (2024). Heterogeneity of consumption responses to income shocks in the presence of nonlinear persistence. *Journal of Econometrics*, 240(2):105449.

Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Blundell, R., Pistaferri, L., and Preston, I. (2008). Consumption inequality and partial insurance. *American Economic Review*, 98(5):1887–1921.

Bonhomme, S. (2021). Teams: Heterogeneity, sorting, and complementarity. *arXiv preprint arXiv:2102.01802*.

Braxton, J. C., Herkenhoff, K., Rothbaum, J., and Schmidt, L. (2024). Changing income risk across the us skill distribution: Evidence from a generalized kalman filter. *Available at SSRN 3983263*.

Chan, J. C. C. and Yu, X. (2022). Fast and accurate variational inference for large bayesian vars with stochastic volatility. *Journal of Economic Dynamics and Control*, 143:104505.

Creal, D. (2012). A survey of sequential monte carlo methods for economics and finance. *Econometric reviews*, 31(3):245–296.

Cremer, C., Li, X., and Duvenaud, D. (2017). Reinterpreting importance-weighted autoencoders. *International Conference on Learning Representations*.

De Nardi, M., Fella, G., and Paz-Pardo, G. (2020). Nonlinear household earnings dynamics, self-insurance, and welfare. *Journal of the European Economic Association*, 18(2):890–926.

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 (methodological)*, 39(1):1–22.

Guvenen, F., Karahan, F., Ozkan, S., and Song, J. (2021). What do data on millions of us workers reveal about lifecycle earnings dynamics? *Econometrica*, 89(5):2303–2339.

Hu, Y. and Schennach, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216.

Jones, M. C. and Pewsey, A. (2009). The sinh-arcsinh normal distribution: Origins and applications. *Biometrika*, 96(4):761–780.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Katsevich, A. and Rigollet, P. (2024). On the approximation accuracy of gaussian variational inference. *The Annals of Statistics*, 52(4):1384–1409.

Kim, Y. and Mnih, A. (2020). Semi-amortized variational inference. *Proceedings of the 37th International Conference on Machine Learning*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *stat*, 1050:1.

Lillard, L. A. and Willis, R. J. (1978). Dynamic aspects of earnings mobility. *Econometrica*, 46(5):985–1012.

Loaiza-Maya, R. and Nibbering, D. (2023). Fast variational bayes methods for multinomial probit models. *Journal of Business & Economic Statistics*, 41(4):1352–1363.

Medina, M. A., Olea, J. L. M., Rush, C., and Velez, A. (2022). On the robustness to misspecification of α-posteriors and their variational approximations. *Journal of Machine Learning Research*, 23(147):1–51.

Meghir, C. and Pistaferri, L. (2004). Income Variance Dynamics and Heterogeneity. *Econometrica*, 72(1):1–32.

Meghir, C. and Pistaferri, L. (2011). Earnings, consumption and life cycle choices. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics, Volume 4B*, pages 773–854. Elsevier.

Mele, A. and Zhu, L. (2023). Approximate variational estimation for a model of network formation. *Review of Economics and Statistics*, 105(1):113–124.

Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. *Proceedings of the 32nd International Conference on Machine Learning*.

Westling, T. and McCormick, T. (2019). Beyond prediction: A framework for inference with variational approximations in mixture models. *Journal of Computational and Graphical Statistics*, 28(4):778–789.

Wu, Y., Burda, Y., Salakhutdinov, R., and Grosse, R. (2016). On the quantitative analysis of decoder-based generative models. *arXiv preprint arXiv:1611.04273*.

A Extensions: Simulation Evidence

A.1 Serially Correlated Transitory Shocks

We consider the model with moving average transitory shocks described in Subsection 6.1. As in the nonlinear model with independent transitory shocks, we consider various forms for the variational posterior density: a Gaussian specification with an unrestricted covariance matrix, a Gaussian with a hidden Markov specification, and a Gaussian with a diagonal covariance matrix (mean-field). In addition, we explore additional structured variational posteriors that incorporate the MA structure, see (40)-(41).

The unrestricted Gaussian posterior provides the most flexible benchmark, as it can capture the full dependence structure implied by the MA(1) shocks. By contrast, the hidden Markov approximation encodes Markovian dependence that is valid under independent transitory shocks. With MA(1) shocks, however, the true posterior does not satisfy this structure. The mean-field Gaussian posterior is even more restrictive, as it rules out all temporal dependence by construction.

We conduct experiments using data simulated from the MA(1) specification described in Subsection 6.1. The moving average coefficient is set to $\zeta = 0.2$, while the remaining elements of the model follow the nonlinear specification used in Section 5, including the distributions of the innovations and the functional forms of $\mu(z_{t-1})$ and $\sigma(z_{t-1})$.

The results, summarized in Figure 8 and Table 5, indicate that, as in the nonlinear model with serially independent transitory shocks, the unrestricted Gaussian variational estimator recovers the mean $\mu(z_{t-1})$ and the volatility $\sigma(z_{t-1})$ quite well (see the first row of Table 5). The hockey-stick shape of $\mu(z_{t-1})$ is well captured, and the estimated $\sigma(z_{t-1})$ function captures some (though not all) of the true nonlinear volatility pattern. The structured variational approach based on (40)-(41), which embeds the dynamic structure of the model to restrict the form of the variational posterior densities, performs similarly (second row). In contrast, both the structured approach based on (misspecified) Markovian variational densities (third row) and the approach based on a Gaussian variational posterior with a diagonal covariance matrix (last row) perform less well. At the same time, none of the variational methods is able to correctly capture the high kurtosis of transitory innovations. Lastly, the unrestricted and (correctly-specified) structured variational approaches estimate a moving average coefficient that is not far from the truth, albeit too low.

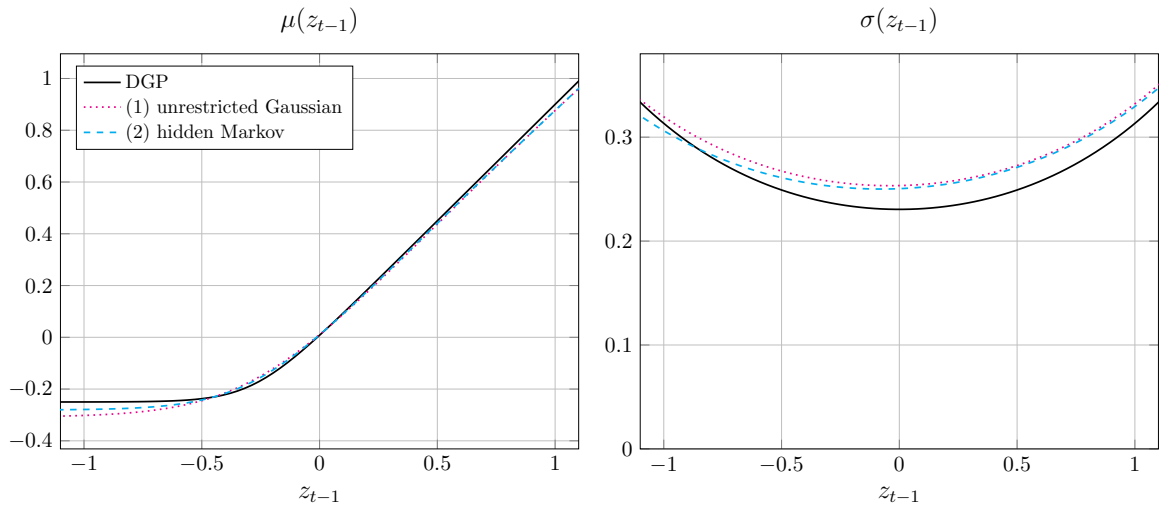

Figure 8: Simulation Results in the Model with MA(1) Transitory Shocks

Note: The figure plots the conditional mean and volatility functions implied by the simulated nonlinear DGP, together with their estimated counterparts obtained under two variational posterior specifications, in the model with MA(1) transitory shocks. Each panel shows the true conditional mean or volatility (solid lines) and their estimated counterparts (dashed lines) across the support of z_{t-1}.

Table 5: Simulation Results in the Model with MA(1) Transitory Shocks

Parameter	α_0	α_1	μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	σ_{z_1}	kurt_{z_1}	σ_ϵ	kurt_ϵ	ζ
DGP	-0.25	0.10	0.00	0.90	0.00	-1.35	0.00	0.35	0.40	3.3	0.16	9.2	0.20
Variational posterior													
(1) unrestricted Gaussian	-0.31	0.24	-0.06	1.02	-0.08	-1.25	0.02	0.29	0.41	3.3	0.12	3.1	0.09
(2) hidden Markov, MA	-0.28	0.17	-0.02	0.95	-0.05	-1.26	0.04	0.27	0.42	3.3	0.13	3.0	0.16
(3) hidden Markov, no MA	-0.34	0.34	-0.13	1.11	-0.13	-1.19	0.07	0.22	0.43	3.3	0.10	3.0	–
(4) diagonal	-0.32	0.14	0.01	0.78	0.07	-1.15	0.16	0.04	0.43	3.3	0.09	3.0	-0.07

Note: The table reports the parameter values used in the simulated nonlinear DGP and the corresponding estimates obtained under four variational posterior specifications. The estimation model allows for nonlinear conditional mean and state-dependent volatility functions, as well as an MA(1) transitory shock, and it is applied to data generated with sinh–arcsinh innovations for z_1 and ϵ_t.

Table 6: Simulations for a Nonlinear Model with Heterogeneous Transitory Variances

Parameter	α_0	α_1	μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	σ_{z_1}	kurt_{z_1}	σ_e	kurt_e	c_0	c_1	c_2
DGP	-0.25	0.10	0.00	0.90	0.00	-1.35	0.00	0.35	0.40	3.3	0.16	9.2	-1.20	0.30	0.30
(1) unrestricted Gaussian	-0.34	0.19	-0.01	0.89	-0.01	-1.28	0.11	0.25	0.41	3.4	0.18	3.2	-1.91	0.10	0.03

Note: The table reports the parameter values used in the simulated nonlinear DGP and the corresponding estimates obtained under an unrestricted Gaussian variational posterior specification. The estimation model allows for nonlinear conditional mean and state-dependent volatility functions, as well as individual-specific transitory variances, and it is applied to data generated with sinh–arcsinh innovations for z_1 and v_t.

A.2 Model with Heterogeneous Transitory Variances

We next consider a nonlinear model with heterogeneity in the variance of transitory shocks, which is a special case of the model introduced in Subsection 6.2:

$$y_t = z_t + e_t, \tag{56}$$

$$z_t = \mu(z_{t-1}) + \sigma(z_{t-1})u_t, \tag{57}$$

$$z_1 \sim f_\alpha, \quad a \,|\, z_1 \sim \mathcal{N}(0,1), \quad u_t \sim \mathcal{N}(0,1), \quad e_t = s(a, z_1)v_t, \quad v_t \,|\, a, z_1 \sim \psi_\gamma, \tag{58}$$

where v_t is distributed according to the sinh-arcsinh family with the scale parameter normalized to one (since the variance of v_t and the magnitude of s are not separately identified). We specify the log-variance as

$$\log s(a, z_1) = c_0 + c_1 z_1 + c_2 a.$$

In Table 6 and Figure 9 we report the estimates based on a simulated sample with $N = 30,0000$ and $T = 10$. We see that the mean and, to a lesser extent, the volatility of z_t given z_{t-1}, are relatively well reproduced. However, Table 6 shows that the features of the distribution of transitory shocks are not well captured. In addition to yielding a low kurtosis, the estimates do not reproduce the patterns of variance heterogeneity present in the model. Regarding computational cost, the nonlinear model with heterogeneous transitory variances is estimated in 4 minutes and 23 seconds.

B Computational Appendix

This appendix describes the implementation of our variational inference estimator, focusing on three components: the Gaussian variational posterior, the quadratic conditionally-Gaussian model, and the sinh–arcsinh error density.

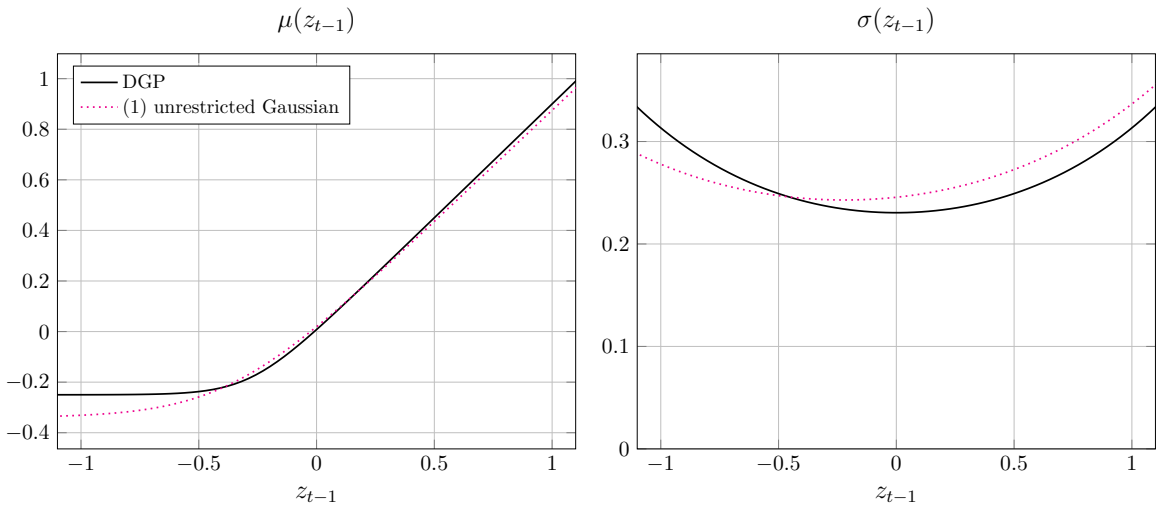

Figure 9: Simulation Results in the Model with Heterogeneous Transitory Variances

Note: The figure plots the conditional mean and volatility functions implied by the simulated nonlinear DGP, together with their estimated counterparts obtained under an unrestricted Gaussian variational posterior specification, in the model with individual-specific transitory variances. Each panel shows the true conditional mean or volatility (solid lines) and their estimated counterparts (dashed lines) across the support of z_{t-1}.

Gaussian Variational Posterior. The variational posterior is a multivariate Gaussian with neural network parameterization. For observation $y \in \mathbb{R}^d$:

$$q(z \mid y) = \mathcal{N}\big(\mu(y),\, L(y)L(y)^\top\big), \tag{59}$$

where $\mu(y) \in \mathbb{R}^{d_z}$ is the mean and $L(y) \in \mathbb{R}^{d_z \times d_z}$ is a lower-triangular Cholesky factor with positive diagonal elements (via softplus). We use the reparameterization trick:

$$z = \mu(y) + L(y) \cdot u, \tag{60}$$

where $u \sim \mathcal{N}(0, I)$.

```
class JointNormalPosterior(nn.Module):
    def __init__(self, dim, dim_latent=None, regularize=1e-3,
                 hidden_dim=32, diagonal=False, sd_clamp=3.0):
        super().__init__()
        self.dim = dim
        self.dim_latent = dim_latent if dim_latent else dim
        self.regularize = regularize
        self.diagonal = diagonal
        self.sd_clamp = sd_clamp

        # Neural network layers
        self.fc = nn.Linear(dim, hidden_dim)
        self.mu = nn.Linear(hidden_dim, self.dim_latent)
```

```
        self.diag_head = nn.Linear(hidden_dim, self.dim_latent)
        if not diagonal:
            n_off = self.dim_latent * (self.dim_latent - 1) // 2
            self.off_head = nn.Linear(hidden_dim, n_off)

    def forward(self, y):
        h = F.relu(self.fc(y))
        mu = self.mu(h)
        mu[..., :self.dim] += y  # Center first latent around observation
        L_diag = F.softplus(self.diag_head(h)) + self.regularize

        if not self.diagonal:
            off_raw = self.off_head(h).clamp(-5.0, 5.0)
            idx_i, idx_j = torch.tril_indices(self.dim_latent,
                                              self.dim_latent, offset=-1)
            L = torch.zeros(y.size(0), self.dim_latent, self.dim_latent)
            L[:, idx_i, idx_j] = off_raw
            L[:, range(self.dim_latent), range(self.dim_latent)] = \
                L_diag.clamp(max=self.sd_clamp)
        else:
            L = torch.diag_embed(L_diag.clamp(max=self.sd_clamp))
        return mu, L

    def draw_and_logprob(self, y, u, logpr_draw=False):
        mu, L = self.forward(y)
        z = mu + torch.matmul(L, u.unsqueeze(-1)).squeeze(-1)
        q = MultivariateNormal(mu, scale_tril=L)
        log_qz = q.log_prob(z) if logpr_draw else -q.entropy()
        return z, log_qz
```

Conditionally-Gaussian Model for $z_1, ..., z_T$. The model $p(z_1, ..., z_T)$ follows a Markov structure:

$$p(z_1, ..., z_T) = p(z_1) \prod_{t=2}^{T} p(z_t \mid z_{t-1}). \tag{61}$$

Each transition is Gaussian:

$$p(z_t \mid z_{t-1}) = \mathcal{N}\big(\mu_\theta(z_{t-1}),\, \sigma^2_\theta(z_{t-1})\big), \tag{62}$$

with polynomial mean and log-standard deviation.

```
class MarkovNormalConditionalPolyPrior(nn.Module):
    """Markov prior with polynomial conditional distributions."""
    def __init__(self, nt, poly_degree=1, regularize=1e-3):
        super().__init__()
        self.nt = nt
        self.regularize = regularize

        # Polynomial functions for conditional mean and log-variance
        self.net_mu = Polynomial(poly_degree)
        self.net_sigma = Polynomial(poly_degree)

        # Initial distribution parameters
        self.mu_z1 = nn.Parameter(torch.zeros(1))
        self.log_sigma_z1 = nn.Parameter(torch.zeros(1))

    def log_prob_z1(self, z1):
```

```
        """Log-probability of initial distribution."""
        return Normal(self.mu_z1, torch.exp(self.log_sigma_z1)).log_prob(z1)

    def log_prob(self, z):
        """Compute log p(z) = log p(z_1) + sum_t log p(z_t | z_{t-1})."""
        logp = self.log_prob_z1(z[:, 0])
        for t in range(1, self.nt):
            z_cur, z_lag = z[:, t:t+1], z[:, t-1:t]
            mu = self.net_mu(z_lag)
            sigma = torch.exp(self.net_sigma(z_lag)) + self.regularize
            logp += Normal(mu, sigma).log_prob(z_cur).squeeze(1)
        return logp
```

Sinh–Arcsinh Error Density. The error density uses a sinh–arcsinh distribution to model heavy tails. For $\widetilde{z} \sim \mathcal{N}(0,1)$:

$$z = \mu + \sigma \cdot \sinh\Big(\delta^{-1} \cdot \big(\operatorname{arcsinh}(\widetilde{z}) + \epsilon\big)\Big), \tag{63}$$

where $\sigma > 0$ is a scale parameter, ϵ controls skewness, and $\delta > 0$ governs tail thickness ($\delta > 1$ implies heavier tails). The inverse is

$$\widetilde{z} = \sinh\Big(\delta \cdot \operatorname{arcsinh}\big((z-\mu)/\sigma\big) - \epsilon\Big), \tag{64}$$

with log-Jacobian:

$$\log\left|\frac{d\widetilde{z}}{dz}\right| = -\log\sigma + \log\delta - \log\cosh\Big(\delta^{-1} \cdot \big(\operatorname{arcsinh}(\widetilde{z}) + \epsilon\big)\Big) + \tfrac{1}{2}\log(1+\widetilde{z}^2). \tag{65}$$

```
class SinhArcsinh(Distribution):
    """Sinh-arcsinh distribution for heavy tails."""
    def __init__(self, loc, scale, skew, tailweight):
        super().__init__()
        self.loc = loc
        self.scale = scale
        self.skew = skew
        self.tailweight = tailweight
        self._base = Normal(0, 1)

    def _inverse(self, y):
        """Inverse transformation: y -> standard normal z."""
        w = (y - self.loc) / self.scale
        return torch.sinh(torch.asinh(w) / self.tailweight - self.skew)

    def log_prob(self, value):
        """Log-probability with Jacobian correction."""
        z = self._inverse(value)
        h = (torch.asinh(z) + self.skew) * self.tailweight
        log_jacobian = -torch.log(self.scale) - torch.log(self.tailweight) \
                       - torch.log(torch.cosh(h)) + 0.5 * torch.log1p(z**2)
        return self._base.log_prob(z) + log_jacobian

class SinhEmission(nn.Module):
    """Decoder with sinh-arcsinh emission distribution."""
    def __init__(self, sigma_eps=0.1, fix_sigma=False):
```

```
        super().__init__()
        self.log_sigma = nn.Parameter(torch.log(sigma_eps * torch.ones(1)),
                                      requires_grad=not fix_sigma)
        self.log_beta = nn.Parameter(torch.zeros(1))  # Tail parameter

    def log_likelihood(self, y, z):
        """Compute log p(y | z) using sinh-arcsinh distribution."""
        dist = SinhArcsinh(loc=torch.zeros(1), scale=torch.exp(self.log_sigma),
                           skew=torch.zeros(1), tailweight=torch.exp(self.log_beta))
        return dist.log_prob(y - z).sum(dim=1)
```

ELBO Objective and Optimization. The complete variational objective maximized via gradient ascent is:

$$\mathcal{E}(\theta, \phi) = \mathbb{E}_{q_\phi(z|y)}\big[\log p_\theta(y \mid z) + \log p_\theta(z) - \log q_\phi(z \mid y)\big]. \tag{66}$$

We estimate the expectation via Monte Carlo sampling using the reparameterization trick. For observations $\{y^{(i)}\}_{i=1}^N$ and S draws per observation:

$$\mathcal{E}(\theta, \phi) \approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{S}\sum_{s=1}^{S}\big[\log p_\theta(y^{(i)} \mid z_s^{(i)}) + \log p_\theta(z_s^{(i)}) - \log q_\phi(z_s^{(i)} \mid y^{(i)})\big], \tag{67}$$

where $z_s^{(i)} \sim q_\phi(\cdot \mid y^{(i)})$ via the reparameterization trick.

```
class VAE(nn.Module):
    """Complete VAE model combining encoder, prior, and decoder."""
    def __init__(self, dim, nt, poly_degree=1, sigma_eps=0.1):
        super().__init__()
        self.encoder = JointNormalPosterior(dim=dim, dim_latent=nt,
                                            hidden_dim=32, regularize=1e-3)
        self.prior = MarkovNormalConditionalPolyPrior(nt=nt,
                                                      poly_degree=poly_degree)
        self.decoder = SinhEmission(sigma_eps=sigma_eps)

    def elbo(self, y, n_samples=1):
        """Compute ELBO estimate with n_samples Monte Carlo samples."""
        batch_size = y.size(0)
        elbo_sum = 0.0

        for _ in range(n_samples):
            # Sample standard normal for reparameterization
            u = torch.randn(batch_size, self.encoder.dim_latent)

            # Encode: sample z ~ q(z|y) and compute log q(z|y)
            z, log_qz = self.encoder.draw_and_logprob(y, u, logpr_draw=True)

            # Decoder: compute log p(y|z)
            log_py_z = self.decoder.log_likelihood(y, z)

            # Prior: compute log p(z)
            log_pz = self.prior.log_prob(z)

            # ELBO = E[log p(y|z) + log p(z) - log q(z|y)]
```

```
            elbo_sum += (log_py_z + log_pz - log_qz).mean()

        return elbo_sum / n_samples

    def forward(self, y):
        """Forward pass returns negative ELBO (loss to minimize)."""
        return -self.elbo(y, n_samples=1)

# Training loop
def train_vae(model, dataloader, num_epochs=100, lr=1e-3):
    optimizer = torch.optim.Adam(model.parameters(), lr=lr)

    for epoch in range(num_epochs):
        epoch_loss = 0.0
        for batch_idx, y_batch in enumerate(dataloader):
            optimizer.zero_grad()

            # Compute negative ELBO (loss)
            loss = model(y_batch)
            loss.backward()
            optimizer.step()

            epoch_loss += loss.item()

        avg_loss = epoch_loss / len(dataloader)
        print(f"Epoch {epoch+1}/{num_epochs}, Loss: {avg_loss:.4f}")

    return model

# Example usage
model = VAE(dim=10, nt=5, poly_degree=2, sigma_eps=0.1)
trained_model = train_vae(model, train_dataloader, num_epochs=100, lr=1e-3)

# Evaluation: compute ELBO with more samples for better estimate
with torch.no_grad():
    test_elbo = model.elbo(test_data, n_samples=100)
    print(f"Test ELBO: {test_elbo:.4f}")
```