GENERATING HEAVY-TAILED FINANCIAL RETURNS VIA ELLIPTICAL VARIATIONAL AUTO-ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Financial risk management often involves learning the return tail distributions, which include both market co-movement and asset-specific extremes. Inspired by the elliptical approximate factor model in the financial literature, we develop an elliptical variational auto-encoder (E-VAE) based on its nonlinear extension, which decomposes financial returns into a nonlinear common factor component and idiosyncratic fluctuations; both follow heavy-tailed distributions with occasional jumps. The elliptical VAE admits a standard evidence lower bound with analytic Kullback–Leibler terms and reparameterized training. In financial applications, the fitted encoder provides posterior estimates of the common factors from historical observations, while the fitted decoder can generate synthetic return vectors for downstream risk assessment and portfolio planning tasks. By exploiting the low-dimensional factor structure, our model enables efficient sampling of high-dimensional return vectors. Experiments on synthetic data and the U.S. and Chinese stock markets demonstrate its ability to capture both the body and tails of return distributions, supporting its use in downstream financial decision-making.

1 INTRODUCTION

Loss control is a central concern in financial risk management (Duffie & Pan, 1997; McNeil et al., 2015). Assessing the loss probability and severity requires a description of the joint return distribution, particularly when losses occur across many assets at once. Modern generative models, such as variational auto-encoders (Kingma & Welling, 2014; Rezende et al., 2014), generative adversarial networks (Goodfellow et al., 2014), and diffusion models (Ho et al., 2020), offer a useful approach. The statistical target of these methods is to learn a distribution of interest from observed data. In the case of financial risk control, once fitted to historical returns, a generative model supplies new synthetic return vectors from which portfolio loss quantiles, expected shortfall, and investment decisions can be evaluated.

Unfortunately, standard generative formulations need not reproduce the heavy tails of financial returns, which are particularly important for risk management (Cont, 2001; Liu et al., 2026). Recently, Cont et al. (2026) proposed generating financial scenarios by matching value-at-risk and expected shortfall for trading strategies via Tail-GAN, which emphasizes risk measures for selected strategies rather than distributional fit across all return levels. Meanwhile, Chen et al. (2025) proposed generating high-dimensional returns using the diffusion factor model. Then, Liu et al. (2026) extended diffusion models to learn from heavy-tailed distributions. However, diffusion sampling is iterative, with a higher computational cost that might matter for financial planning when many simulated returns are required. Our aim is, therefore, to *obtain a computationally efficient sampler of the joint return distribution that captures both ordinary fluctuations and heavy-tailed behavior*.

The intrinsic low-dimensional factor structure of financial returns offers a natural way to pursue this aim. The approximate factor model $y_t = AF_t + e_t$ represents common movements through a low-dimensional factor vector F_t, a loading matrix A, and idiosyncratic returns e_t (Bai, 2003). This reduces the problem of modeling dependence among many assets to learning the common factors and their effects on returns. The linear factor model formulation has been extended to accommodate both heavy tails and nonlinear dependence. Elliptical factor models allow for heavy-tailed variation while retaining a tractable factor structure (Fan et al., 2018; He et al., 2022). Meanwhile, nonlinear factor models replace the linear map AF_t with an unknown function $f(F_t)$, which can be estimated using

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Conceptual correspondence between the financial factor model and a VAE.

Financial factor model	VAE formulation
Common factors F_t	Latent variables z_t
Factor distribution	Prior $p_\psi(z_t)$
Factor-to-return function f	Decoder mean f_θ
Returns conditional on factors	Decoder likelihood $p_\theta(y_t \mid z_t)$
Factor inference from observed returns	Encoder posterior $q_\phi(\cdot \mid y_t)$

deep autoencoders (Xiu & Shen, 2024). These extensions address different features of returns and can be considered together. As such, in this work, we introduce a *nonlinear elliptical factor model* in which returns depend nonlinearly on common factors, and both the factors and idiosyncratic returns can be heavy-tailed. For $i \in [N]$ and $t \in [T]$, we write

$$y_t = f(F_t) + e_t, \quad F_t = R_{f,t}U_{f,t}, \quad e_{it} = s_i R_{i,t} U_{i,t}. \tag{1}$$

Here $F_t \in \mathbb{R}^k$ with $k \ll N$, $f : \mathbb{R}^k \to \mathbb{R}^N$ is the unknown factor-to-return function, and $s_i > 0$ is an asset-specific scale. We normalize the factor scatter to the identity since any non-singular scatter transformation can be absorbed into the unknown function f. We assume independence between the factor variables and the idiosyncratic variables. The directions $U_{f,t}$ and $U_{i,t}$ are uniform on $\mathbb{S}^{k-1}$ and $\{-1, 1\}$, respectively. The radii $R_{f,t}$ and $R_{i,t}$, independent of the directions, determine the magnitudes of common and asset-specific fluctuations and can be heavy-tailed. The factors and each idiosyncratic return thus belong to the elliptical family, which includes Gaussian and Student-t distributions as special cases.

Learning this model from observed returns requires estimating the function f and the distributions of the unobserved factors and residuals. Interestingly, the conceptual correspondence between factor models and VAEs, summarized in table 1, suggests a natural solution through an *elliptical variational auto-encoder* (E-VAE). In particular, the factor distribution serves as the prior, while the conditional return distribution defines the decoder likelihood, with a neural network f_θ approximating f. The encoder approximates the posterior distribution of the latent variables given observed returns.

In financial applications, jump models distinguish ordinary fluctuations from abrupt moves, with empirical evidence for both systematic and idiosyncratic jumps (Merton, 1976; Bollerslev et al., 2013). This motivates separate Student-t mixtures for common factors and asset-specific residuals. Earlier VAEs have used Student-t priors or observation models (Abiri & Ohlsson, 2019; Takahashi et al., 2018), while t^3-VAE and Pareto-VAE use power-divergence objectives (Kim et al., 2024; Cho et al., 2026). Meanwhile, our mixture modeling distinguishes ordinary and jump states for each source of financial risk, with separate frequencies, scales, and tail indices. Removing the jump components recovers Student-t laws, with Gaussian laws as the limiting cases. In E-VAE, the independence of Gaussian radii and directions leads to a scale-mixture representation with closed-form Gaussian, Gamma, and Bernoulli KL terms in a standard ELBO. This construction also accommodates other Gaussian scale mixtures with tractable scale KL terms, and the proposed jump mixture is mainly used to model financial return behavior. Generation requires one decoder evaluation, together with factor and residual draws.

The paper makes three contributions.

1. For accurate and efficient financial distribution learning, we introduce a nonlinear elliptical factor model and its E-VAE estimator, with separate common and private heavy tails and jumps. Gaussian, Student-t, and jump-mixture specifications belong to the same family, linking classical low-dimensional financial factor structures to flexible return generation.
2. We derive a standard ELBO with closed-form KL terms for the augmented latent variables while integrating out private states analytically. Sampling requires one decoder pass. On the U.S. panel, E-VAE-CP generates 4,000 return vectors in 0.033 seconds on an Apple M5 CPU, offering a promising and computationally inexpensive solution for building financial simulation environments and evaluating investment strategies.
3. Experiments on synthetic data and the U.S. and Chinese stock markets show that E-VAE captures tail risk while retaining a close fit to the body. It achieves a better overall distributional fit than Tail-GAN and Tail-Diffusion. Our empirical results suggest that modeling common jumps brings

greater portfolio benefits in the U.S., whereas modeling private jumps is more beneficial in China. These findings highlight the importance of distinguishing market-wide and asset-specific sources of tail risk in portfolio management.

2 Elliptical Variational Auto-Encoders

In this section, we present E-VAE as an estimator of the proposed nonlinear elliptical factor model. Throughout, f denotes the population function in equation (1), whereas f_θ denotes the neural network used to approximate it. We begin with a brief review of the Gaussian VAE.

2.1 Preliminaries on Gaussian VAE

The Gaussian case of equation (1) takes $R_{f,t} \sim \chi_k$ and $R_{i,t} \sim \chi_1$. With independent residuals, replacing f with a neural network gives the generative model

$$z_t \sim \mathcal{N}(0, I_k), \qquad y_t = f_\theta(z_t) + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, D), \quad D = \mathrm{diag}(s_1^2, \ldots, s_N^2), \tag{2}$$

where z_t and ϵ_t are independent. An affine decoder $f_\theta(z) = c + Az$ gives $y_t \sim \mathcal{N}(c, AA^\top + D)$, the Gaussian factor analysis model. If $D = s^2 I_N$, this is precisely the probabilistic PCA model (Tipping & Bishop, 1999; Lucas et al., 2019). A nonlinear decoder is able to extend the factor-to-return relation while keeping the Gaussian prior and residual law.

To estimate θ, which includes the network weights and residual scales, the likelihood integrates out the unobserved factors by $p_\theta(y_t) = \int p_\theta(y_t \mid z_t) p(z_t)\, \mathrm{d}z_t$. This integral is generally intractable for a nonlinear f_θ. A Gaussian VAE approximates the factor posterior by $q_\phi(z_t \mid y_t) = \mathcal{N}(\mu_\phi(y_t), \mathrm{diag}\{\sigma_\phi^2(y_t)\})$ and estimates θ and ϕ by maximizing the evidence lower bound (ELBO)

$$\mathcal{L}_{\mathrm{G}}(y_t) = \mathbb{E}_{q_\phi(z_t \mid y_t)}[\log p_\theta(y_t \mid z_t)] - \mathrm{KL}(q_\phi(z_t \mid y_t) \,\|\, \mathcal{N}(0, I_k)) \le \log p_\theta(y_t). \tag{3}$$

With fixed s_i, maximizing the expected log likelihood minimizes a weighted squared reconstruction error. When s_i are estimated, the likelihood also includes their log normalizers. The Gaussian KL is analytic, and the likelihood expectation is estimated using reparameterized draws $z_t = \mu_\phi(y_t) + \sigma_\phi(y_t) \odot \eta_t$, where $\eta_t \sim \mathcal{N}(0, I_k)$. Differentiating these draws gives stochastic gradients for the encoder and decoder (Kingma & Welling, 2014; Rezende et al., 2014). The complete objective is given in section A.

Although a nonlinear decoder can fit a non-Gaussian return distribution, it cannot, by itself, remove the light-tail restriction of Gaussian generation. The following result concerns the distribution generated by a fitted Gaussian VAE, with its parameters held fixed. Its Lipschitz condition holds for finite feedforward networks with finite weights and Lipschitz activations such as ReLU, tanh, and softmax.

Theorem 1 (Tails of a fitted Gaussian VAE)**.** *Let y_t be generated by equation (2), with fixed parameters and a globally L_θ-Lipschitz decoder f_θ in the Euclidean norm. For any $w \in \mathbb{R}^N$, define $X_w = w^\top(y_t - \mathbb{E}y_t)$. Then*

$$\mathbb{E}e^{\lambda X_w} \le \exp\left\{ \frac{\lambda^2}{2} \left(L_\theta^2 \|w\|^2 + w^\top D w \right) \right\}, \qquad \lambda \in \mathbb{R}.$$

Thus, every centered portfolio return generated by the fitted model is sub-Gaussian, and

$$\lim_{x \to \infty} x^\alpha \mathbb{P}(|X_w| > x) = 0 \qquad \textit{for every } \alpha > 0.$$

In particular, the fitted model cannot reproduce a target portfolio tail with $\mathbb{P}(|X| > x) \sim c_0 x^{-\alpha}$ for $c_0, \alpha > 0$.

The proof is given in section A.2. This light-tail restriction motivates us to introduce heavy-tailed distributions for the factors and residuals.

2.2 Elliptical factors and asset-specific heavy tails

For mathematical and computational convenience, we implement the elliptical model through independent Gaussian scale mixtures (Andrews & Mallows, 1974):

$$\begin{aligned} z_t &= \sqrt{V_{f,t}}\, G_t, \qquad G_t \sim \mathcal{N}(0, I_k), \qquad E_{it} \sim \mathcal{N}(0,1), \\ y_{it} &= f_{\theta i}(z_t) + s_i \sqrt{V_{i,t}}\, E_{it}, \quad i = 1, \ldots, N. \end{aligned} \tag{4}$$

The Gaussian variables and positive variance multipliers are mutually independent. Indeed, a standard Gaussian has an independent chi radius and uniform direction, so the radii in equation (1) are $R_{f,t} = \sqrt{V_{f,t}}Q_{f,t}$ and $R_{i,t} = \sqrt{V_{i,t}}Q_{i,t}$, with $Q_{f,t} \sim \chi_k$ and $Q_{i,t} \sim \chi_1$. This gives a scale-mixture subclass of the elliptical family while retaining Gaussian conditional distributions; section B derives the representation. The common multiplier $V_{f,t}$ changes factor magnitudes, whose effects across assets are determined by f_θ. Each private multiplier $V_{i,t}$ acts on one asset's residual. Their distributions can therefore describe different common and private tails. Note that setting all multipliers to one recovers equation (2).

For an affine function $f(F_t) = AF_t$, conditioning on the scale multipliers gives a covariance with a low-rank common component and a diagonal private component. Replacing the independent multipliers with one shared multiplier gives a joint elliptical return distribution, as in classical elliptical factor models (Fan et al., 2018; He et al., 2022). Separate multipliers allow common and private tails to differ. In section 3.2.4, we examine the implications for extreme return directions.

2.3 Common and private jump distributions

To model the potential jump behavior of financial returns (Merton, 1976; Bollerslev et al., 2013), we use a two-component Student-t mixture for the factors to describe ordinary fluctuations and occasional jumps. Let

$$B_{f,t} \sim \text{Bernoulli}(\pi_f), \qquad \Lambda_{f,t} \,|\, B_{f,t} = b \sim \text{Gamma}\Big(\frac{\nu_{f,b}}{2}, \frac{\nu_{f,b}}{2}\Big),$$
$$V_{f,t} = \frac{a^2_{f,B_{f,t}}}{\Lambda_{f,t}}, \qquad a_{f,0} = 1, \quad a_{f,1} > 1.$$

We use the shape–rate convention for Gamma distributions. The state $B_{f,t} = 1$ represents a common jump over the observation interval, with probability π_f and relative scale $a_{f,1}$. Within each state, the precision $\Lambda_{f,t}$ generates a Student-t law whose degrees of freedom $\nu_{f,b}$ determine its tail decay. Integrating out the state and precision gives the factor prior

$$p_\psi(z_t) = (1-\pi_f)t^{(k)}_{\nu_{f,0}}(z_t; 0, I_k) + \pi_f t^{(k)}_{\nu_{f,1}}(z_t; 0, a^2_{f,1} I_k). \tag{5}$$

Here, the last argument of $t^{(k)}_\nu$ is the scatter matrix, and ψ collects the factor distribution parameters. Both components have the same uniform direction and differ only in their radial distributions. Private jumps are specified in the same way, independently across assets and independently of the factors:

$$B_{i,t} \sim \text{Bernoulli}(\rho_i), \qquad \Lambda_{i,t} \,|\, B_{i,t} = b \sim \text{Gamma}\Big(\frac{\nu_{i,b}}{2}, \frac{\nu_{i,b}}{2}\Big),$$
$$V_{i,t} = \frac{a^2_{i,B_{i,t}}}{\Lambda_{i,t}}, \qquad a_{i,0} = 1, \quad a_{i,1} > 1.$$

The probability ρ_i controls the frequency of private jumps, while s_i and $a_{i,1}s_i$ set the ordinary and jump scales. Integrating out the private variables gives the decoder likelihood

$$p_\theta(y_t \,|\, z_t) = \prod_{i=1}^{N} \big[(1-\rho_i)t_{\nu_{i,0}}(y_{it}; f_{\theta i}(z_t), s_i) + \rho_i t_{\nu_{i,1}}(y_{it}; f_{\theta i}(z_t), a_{i,1}s_i)\big]. \tag{6}$$

The last argument of a univariate t density is its scale, and θ includes the residual distribution parameters as well as the decoder weights. We take active jump probabilities in $(0, 1/2)$ and set $\nu_{f,1} \le \nu_{f,0}$, $\nu_{i,1} \le \nu_{i,0}$, so the jumps are less frequent, have larger scales, and have equally heavy or heavier tails. For finite-covariance specifications, all active degrees of freedom exceed 2, and f_θ has at most linear growth. The density, variance, and tail calculations are given in section B.

2.4 Variational estimation

The private variables have been integrated out in equation (6), leaving the common Gaussian variable G_t, state $B_{f,t}$, and precision $\Lambda_{f,t}$ for variational inference. We use

$$q_\phi(g, \lambda, b \,|\, y_t) = q_\phi(g \,|\, y_t) q_\phi(b \,|\, y_t) q_\phi(\lambda \,|\, y_t, b).$$

The encoder distributions are

$$q_\phi(G_t \mid y_t) = \mathcal{N}\big(\mu_\phi(y_t), \operatorname{diag}\{\sigma^2_\phi(y_t)\}\big), \qquad q_\phi(B_{f,t} = 1 \mid y_t) = r_\phi(y_t),$$
$$q_\phi(\Lambda_{f,t} \mid y_t, B_{f,t} = b) = \operatorname{Gamma}(\alpha_{\phi,b}(y_t), \beta_{\phi,b}(y_t)). \tag{7}$$

The Gaussian posterior is shared across the two states, while each state has its own Gamma posterior. The encoder outputs positive scales, shapes, and rates, and a probability $r_\phi(y_t) \in (0,1)$. Since $z_t = a_{f,b}\lambda^{-1/2} g$, the joint density is

$$p_{\theta,\psi}(y_t, g, \lambda, b) = p_\theta(y_t \mid a_{f,b}\lambda^{-1/2} g)\, \mathcal{N}(g; 0, I_k)\, p_\psi(b)\, p_\psi(\lambda \mid b).$$

Applying Jensen's inequality and separating the two common states gives

$$\begin{aligned}
\mathcal{L}(y_t) = {} & (1 - r_\phi(y_t))\mathbb{E}_{\substack{G_t \sim q_\phi(G_t \mid y_t) \\ \Lambda_{f,t} \sim q_\phi(\Lambda_{f,t} \mid y_t, B_{f,t}=0)}} \log p_\theta\Big(y_t \mid \Lambda_{f,t}^{-1/2} G_t\Big) \\
& + r_\phi(y_t)\mathbb{E}_{\substack{G_t \sim q_\phi(G_t \mid y_t) \\ \Lambda_{f,t} \sim q_\phi(\Lambda_{f,t} \mid y_t, B_{f,t}=1)}} \log p_\theta\Big(y_t \mid a_{f,1}\Lambda_{f,t}^{-1/2} G_t\Big) \\
& - \frac{1}{2}\sum_{j=1}^{k}\left[\mu_{\phi j}(y_t)^2 + \sigma_{\phi j}(y_t)^2 - 1 - \log \sigma_{\phi j}(y_t)^2\right] \\
& - (1 - r_\phi(y_t))\operatorname{KL}\Big(\operatorname{Gamma}(\alpha_{\phi,0}(y_t), \beta_{\phi,0}(y_t)) \,\Big\|\, \operatorname{Gamma}\Big(\frac{\nu_{f,0}}{2}, \frac{\nu_{f,0}}{2}\Big)\Big) \\
& - r_\phi(y_t)\operatorname{KL}\Big(\operatorname{Gamma}(\alpha_{\phi,1}(y_t), \beta_{\phi,1}(y_t)) \,\Big\|\, \operatorname{Gamma}\Big(\frac{\nu_{f,1}}{2}, \frac{\nu_{f,1}}{2}\Big)\Big) \\
& - (1 - r_\phi(y_t)) \log \frac{1 - r_\phi(y_t)}{1 - \pi_f} - r_\phi(y_t) \log \frac{r_\phi(y_t)}{\pi_f} \;\le\; \log p_{\theta,\psi}(y_t).
\end{aligned} \tag{8}$$

The shared Gaussian posterior contributes one KL term; each Gamma KL is weighted by its state's posterior probability. Here, we share the Gaussian variational factor across the ordinary and jump states mainly for computational simplicity. This restriction can be relaxed to state-dependent Gaussian factors $q_\phi(G_t \mid y_t, B_{f,t} = b)$, while retaining analytic KL terms in the augmented ELBO. All KL terms are analytic. Private jumps enter through the integrated mixture likelihood. Section C derives the bound and the KL expressions. The same objective covers the reduced models. Setting $\pi_f = 0$ requires $r_\phi(y_t) = 0$ for finite KL, so the common jump branch and Bernoulli KL are absent. Setting $\rho_i = 0$ removes the private jump component from asset i's likelihood. After removing the jumps and the heavy-tailed components, the objective is exactly equation (3). The reduced expressions are given in section D.

2.5 Training and Generation

We estimate the model by maximizing $T^{-1}\sum_{t=1}^{T}\mathcal{L}(y_t)$ with minibatch stochastic gradients. For serially dependent returns, this is a marginal fitting criterion for the one-period return distribution.

The likelihood expectations use reparameterized Gaussian draws $G_t = \mu_\phi(y_t) + \sigma_\phi(y_t) \odot \eta_t$, where $\eta_t \sim \mathcal{N}(0, I_k)$, and Gamma draws with implicit reparameterization gradients (Figurnov et al., 2018). Both common states are evaluated and weighted by their posterior probabilities, so training does not require sampling a Bernoulli state. Private states are marginalized analytically, with likelihood evaluation scaling linearly with respect to the number of assets. The encoder and decoder parameters, together with the asset-specific residual scales, are jointly optimized. Mixture probabilities, relative scales, and degrees of freedom are treated as hyperparameters and held fixed during training. Implementation details are provided in section E.

To generate returns, sample the common state, precision, and Gaussian vector from their priors and set $z_t = a_{f,B_{f,t}}\Lambda_{f,t}^{-1/2} G_t$. We can evaluate $f_\theta(z_t)$ once, then draw the private variables independently to obtain $y_{it} = f_{\theta i}(z_t) + s_i a_{i,B_{i,t}}\Lambda_{i,t}^{-1/2} E_{it}$. Details concerning historical posterior summaries using the trained encoder are given in section F.

3 Numerical Experiments

Our experiments examine whether heavy-tailed factors and residuals improve the generation of extremes while retaining the VAE's fit to the ordinary return distribution. We study three synthetic

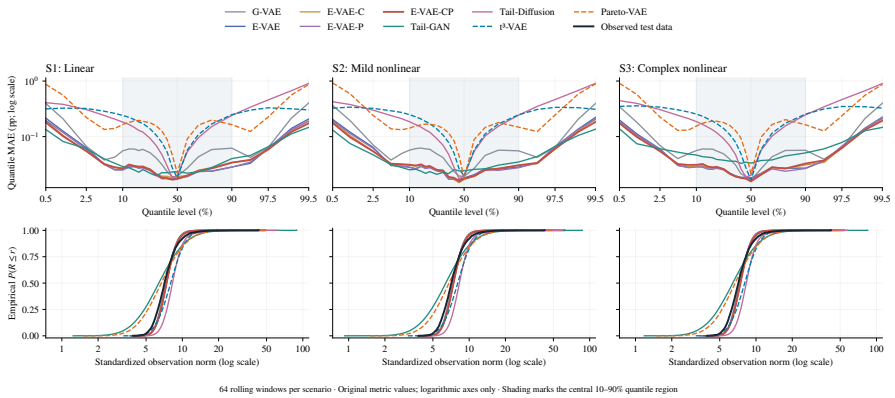

Figure 1: Synthetic designs S1–S3. Top: QMAE (pp), with log vertical and logit horizontal axes; shading marks the 10–90% quantiles. Bottom: standardized observation-norm CDFs on log horizontal axes; black denotes held-out data.

designs and two equity markets. Then, we evaluate the generated returns in a downstream portfolio management task. Four members of the proposed family distinguish the role of jumps: E-VAE without jumps, E-VAE-C with common jumps, E-VAE-P with private jumps, and E-VAE-CP with both. We compare them with Gaussian VAE (G-VAE), Tail-GAN (Cont et al., 2026), Tail-Diffusion (Liu et al., 2026), and two heavy-tailed VAE baselines, t^3-VAE (Kim et al., 2024) and Pareto-VAE (Cho et al., 2026), using the same training, validation, and evaluation windows. Section H gives the designs and evaluation measures. An anonymous implementation of E-VAE is currently available at `https://anonymous.4open.science/r/E_VAE/`.

3.1 Synthetic Data

Each design contains $T = 6{,}000$ independent observations of $N = 60$ assets driven by $k = 5$ factors. The factors and residuals have the common and private Student-t mixtures in section 2.3. The designs differ only in the factor-to-return function: S1 is linear, S2 adds a bounded nonlinear response, and S3 adds another nonlinear layer and an even response component. Loading matrices are drawn once and fixed; an independent pilot sample sets the marginal return scales. The complete specifications are in section H.1.

We measure quantile errors by $\mathrm{QMAE}(\alpha) = 100N^{-1}\sum_i |\widehat{q}_i^{\mathrm{gen}}(\alpha) - \widehat{q}_i^{\mathrm{obs}}(\alpha)|$ in return percentage points. Across all three designs in figure 1, E-VAE-CP combines small tail errors with a close fit across interior quantiles. Although Tail-GAN achieves the lowest errors at the most extreme quantiles, this accuracy does not extend to the distribution of return magnitudes: the norm CDFs indicate an excess of small-magnitude samples. Tail-Diffusion, instead, shifts mass toward larger magnitudes. E-VAE-CP follows the held-out CDF through its main transition, showing that accurate tail generation can be achieved together with a close fit to the body. This pattern persists as the factor-to-return map becomes more nonlinear.

3.2 U.S. and Chinese Stock Markets

The collected U.S. and Chinese panels contain 158 and 95 assets, respectively. Each model is fitted on a rolling five-year history (the first 90% for training and the remaining 10% for validation) and tested in the next quarter. The history then advances by one quarter, dropping the oldest quarter and adding the newest. This yields 64 U.S. and 62 Chinese evaluation windows, spanning 4,008 and 3,752 trading days, respectively. Within each market, all nine models share observation dates and fitting/validation splits. Details appear in section H.3.

3.2.1 Distributional fit

Figure 2 compares lower-tail quantile accuracy and overall distributional fit in both markets. At 2.5%, E-VAE-C ranks second in the U.S. and first in China, while E-VAE-CP performs best at 5% in

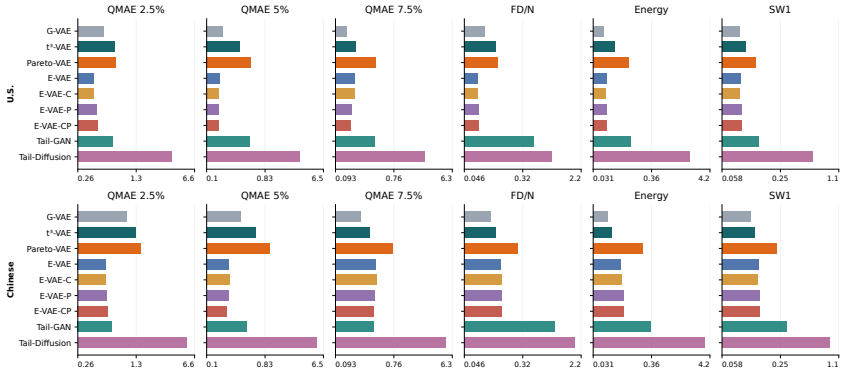

Figure 2: Held-out fit in the U.S. (top) and Chinese stock markets (bottom). Columns 1–3 report QMAE (return pp); columns 4–6 report FD/N, Energy, and SW1. Numeric axes are logarithmic, with matching ranges across markets. Smaller is better.

Table 2: CB portfolios over the full evaluation period. SR: annualized Sharpe ratio; CR: cumulative net excess return; Mean/yr: annualized mean net excess return; MDD: maximum drawdown; ES95: daily expected shortfall; Cost/yr: annualized trading cost. All entries except SR are percentages. Bold uses unrounded results.

	U.S. stock market						Chinese stock market					
Model	SR	CR	Mean/yr	MDD	ES95	Cost/yr	SR	CR	Mean/yr	MDD	ES95	Cost/yr
G-VAE	.555	225.7	8.64	41.65	2.35	1.22	.598	591.0	17.06	71.61	4.47	1.54
t^3-VAE	.631	313.6	**10.25**	37.24	2.46	1.37	.624	664.7	17.71	70.12	4.47	1.66
Pareto-VAE	.580	274.5	9.70	**36.42**	2.49	**1.17**	.652	779.0	18.76	70.86	4.63	1.76
E-VAE	.613	268.3	9.37	38.75	**2.28**	1.22	.710	970.0	19.84	**66.84**	**4.31**	1.46
E-VAE-C	**.671**	**323.4**	10.24	38.71	2.30	1.23	.634	688.0	17.84	68.10	4.37	**1.44**
E-VAE-P	.658	314.1	10.12	37.11	2.30	1.22	**.729**	**1064.1**	**20.41**	69.09	4.31	1.47
E-VAE-CP	.654	308.3	10.02	39.79	2.32	1.19	.610	618.8	17.25	68.39	4.40	1.44
Tail-GAN	.455	232.4	9.96	57.98	3.37	1.92	.610	685.0	18.44	73.10	4.89	2.06
Tail-Diffusion	.429	179.0	8.35	54.12	3.01	1.51	.678	854.0	19.16	69.34	4.46	1.71

both markets. Its errors are 0.161 and 0.214 percentage points in the U.S. and China, respectively, compared with 0.481 and 0.432 for Tail-GAN. At the same 5% quantile, t^3-VAE records errors of 0.337 and 0.590 percentage points, while Pareto-VAE records 0.497 and 0.966, respectively. At 7.5%, E-VAE-CP ranks second to G-VAE in the U.S. and third in China, behind G-VAE and t^3-VAE.

These gains in tail accuracy are accompanied by a strong overall distributional fit. Every E-VAE variant achieves lower FD/N, energy discrepancy, and sliced Wasserstein distance (SW1) than both Tail-GAN and Tail-Diffusion in both markets. The E-VAE variants also outperform both t^3-VAE and Pareto-VAE on these three measures in the U.S. In China, they outperform Pareto-VAE, whereas t^3-VAE achieves lower discrepancies than the E-VAE variants. G-VAE also provides a strong overall fit, while all E-VAE variants improve upon its quantile accuracy at 2.5% and 5% in both markets. Taken together, the results support the elliptical specification as a competitive approach to representing extreme losses while retaining a close fit to the broader return distribution in financial applications.

3.2.2 Downside-aware portfolio evaluation

Following the CVaR-constrained portfolio framework of Krokhmal et al. (2002), we evaluate investment performance using a tail-aware CVaR-budget (CB) allocation rule. Using 1,000 model-generated return scenarios per window, we maximize expected return minus a turnover penalty under a 95% CVaR budget set 10% above its estimated minimum. All models use the same rule, with long-only, fully invested portfolios and a 5% asset cap. Quarterly target weights are restored daily at a cost of 0.002 times full ℓ_1 turnover. Details appear in section H.4.

In the U.S., E-VAE-C has the highest full-period SR and CR, at 0.671 and 323.4%, compared with 0.613 and 268.3% for no-jump E-VAE (table 2). In China, E-VAE-P leads at 0.729 and 1064.1%, compared with 0.710 and 970.0%. Both winners also have smaller drawdowns and a smaller expected

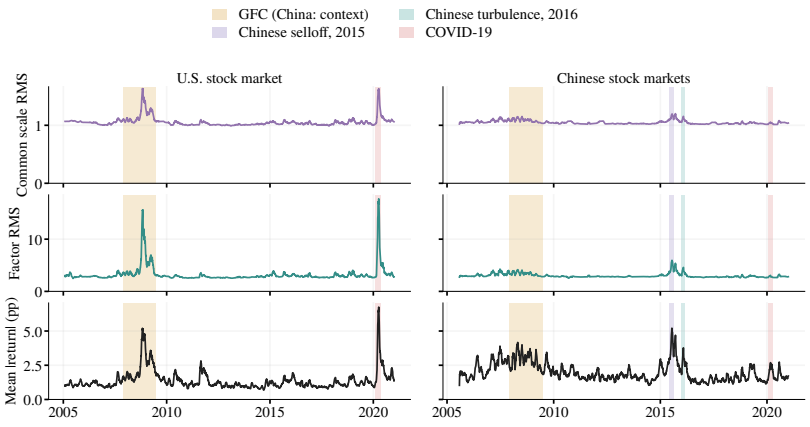

Figure 3: Historical E-VAE-CP posterior estimates and observed return magnitudes. Rows show common-scale RMS (purple), factor RMS (teal), and cross-sectional mean absolute return in pp (black), all with 21-observation trailing means. Bands mark market episodes; each row shares its y-axis range across markets. Details are in section H.5.

shortfall than Tail-GAN. Allowing common jumps is therefore more useful for this portfolio rule in the U.S., while private jumps are more useful in China.

The proposed method has a low computational cost in the portfolio application, where rolling-window evaluation requires repeated model fitting, followed by the generation of many return scenarios for investment decisions. On the U.S. panel, E-VAE-CP takes an average of 30.81 seconds per window for fitting and generation on an Apple M5 CPU, and generates 4,000 return vectors in 0.033 seconds. This combination is useful for financial simulation environments and strategy evaluation, where the return model must be updated as new observations become available and sampled repeatedly for downstream decisions. For reference, our Tail-GAN implementation takes 1,640.76 seconds per window for training, including validation and checkpoint saving, on an NVIDIA GeForce RTX 5060 Ti GPU. Section H.6 reports the hardware configurations and timing details.

3.2.3 Historical posterior estimates

Meanwhile, the fitted encoder is able to provide a complementary view of these market differences by summarizing common variation over time. Figure 3 compares the posterior common-scale RMS, $\{\mathbb{E}_{q_\phi}[V_{f,t} \mid y_t]\}^{1/2}$, and factor RMS, $\{\mathbb{E}_{q_\phi}[\|z_t\|^2 \mid y_t]\}^{1/2}$, for E-VAE-CP with the observed cross-sectional mean absolute return. The former summarizes the scale assigned to common shocks, while the latter measures the overall magnitude of the inferred factors. Together, they show how the model represents common fluctuations during periods of market stress.

Both posterior summaries rise around the global financial crisis and COVID-19 in the U.S., accompanying the increase in observed return magnitudes. In China, factor RMS rises around the 2015 selloff, while the common scale varies less. These historical summaries give a financial interpretation to the latent variables without requiring market episodes to be labeled during training. Derivations and implementation details are provided in sections F and H.5.

3.2.4 Beyond a fixed linear factor subspace

Finally, we examine the geometry of ordinary and extreme returns. A linear factor model accommodates heavy-tailed common factors and asset-specific shocks but transmits all common movements through the same loading matrix. To characterize this restriction, consider the affine population model

$$y_t = \sqrt{V_{f,t}}\, A G_t + e_t, \qquad e_{it} = s_i \sqrt{V_{i,t}} E_{it}. \tag{9}$$

The limiting distribution of $(y_t)/\|y_t\|$ conditional on a large norm describes the directions of extreme returns. However, our empirical analysis indicates that angular variation remains concentrated in low-dimensional subspaces in both the body and the tail, but the dominant subspaces differ (Figure 4, left).

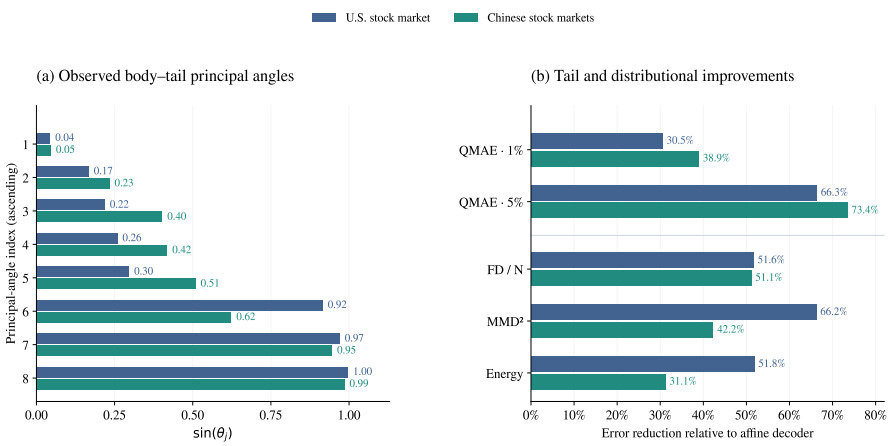

Figure 4: Left: body-tail principal-angle sines (eight directions). Right: nonlinear E-VAE-CP error reductions relative to its affine counterpart. Both panels use full evaluation periods.

Theorem 2 (A fixed common subspace in covariance and tails)**.** *In equation (9), let $A \neq 0$, $s_i > 0$, and assume the independence conditions in equation (4). Suppose $\mathbb{P}(\sqrt{V_{f,t}} > x) \sim c_f x^{-\alpha_f}$ and $\mathbb{P}(\sqrt{V_{i,t}} > x) \sim c_i x^{-\alpha_i}$, with positive constants and exponents. If second moments are finite and $\alpha_f \leq \min_i \alpha_i$, then*

$$\mathrm{Cov}(y_t) = \mathbb{E}[V_{f,t}]AA^\top + D, \qquad M_{\mathrm{tail}} = \omega M_f + (1-\omega)D_e,$$

where D and D_e are diagonal, $\omega > 0$, and $\mathrm{range}\big(\mathbb{E}[V_{f,t}]AA^\top\big) = \mathrm{range}(M_f) = \mathrm{range}(A)$. Thus, the common covariance and common tail angular components share the same loading subspace. In particular, if $\alpha_f < \min_i \alpha_i$, then $\omega = 1$ and $\mathrm{range}(M_{\mathrm{tail}}) = \mathrm{range}(A)$.

More generally, the tail decomposition holds without finite second moments, with $\omega \in [0,1]$ and $\omega > 0$ exactly when $\alpha_f \leq \min_i \alpha_i$. The definitions of M_{tail}, M_f, and D_e are given in Appendix G.5.

The nonlinear decoder in E-VAE provides the flexibility for heavy-tailed common movements, as it allows its image under f_θ to depart from a fixed linear loading space. Consequently, typical and extreme factor values can generate different dominant return directions. To assess the practical value of this flexibility, we compare E-VAE-CP with its affine counterpart while holding the encoder, factor dimension, radial distributions, data splits, and training settings fixed. Over the full evaluation periods, nonlinear decoding reduces the reported tail-quantile errors and joint distributional discrepancies in both markets (table 5). Together, the geometric analysis motivates relaxing the fixed loading restriction, while the paired ablation demonstrates the resulting improvement in tail and overall distributional fit.

4 Discussion

In this work, we propose E-VAE that connects nonlinear financial factor models with variational generation. Independent common and private scales introduce heavy tails and jumps within a tractable likelihood, with analytic KL terms and a single decoder evaluation for sampling. Experiments on synthetic data and the U.S. and Chinese stock markets show that this construction improves tail modeling while retaining a close fit to the body of the return distribution. The portfolio results further reveal distinct benefits from common and private jumps across markets.

A natural next step is to use the learned return distribution for sequential portfolio planning and reinforcement learning. Conditioning E-VAE on market information and recent returns would support state-dependent generation. Coupled with portfolio transitions, transaction costs, and risk-sensitive rewards, this would provide an environment model for learning investment policies from simulated trajectories, following model-based reinforcement learning (Sutton, 1990; Hafner et al., 2019; 2020).

REFERENCES

Najmeh Abiri and Mattias Ohlsson. Variational auto-encoders with student's t-prior. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 415–420, 2019.

David F. Andrews and Colin L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):99–102, 1974. doi: 10.1111/j.2517-6161.1974.tb00989.x.

Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171, 2003. doi: 10.1111/1468-0262.00392.

Bojan Basrak, Richard A. Davis, and Thomas Mikosch. Regular variation of GARCH processes. *Stochastic Processes and their Applications*, 99(1):95–115, 2002. doi: 10.1016/S0304-4149(01)00156-9.

Tim Bollerslev, Viktor Todorov, and Sophia Zhengzi Li. Jump tails, extreme dependencies, and the distribution of stock returns. *Journal of Econometrics*, 172(2):307–324, 2013. doi: 10.1016/j.jeconom.2012.08.014.

Leo Breiman. On some limit theorems similar to the arc-sin law. *Theory of Probability & Its Applications*, 10(2):323–331, 1965. doi: 10.1137/1110037.

Stamatis Cambanis, Steel Huang, and Gordon Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981. doi: 10.1016/0047-259X(81)90082-8.

Minshuo Chen, Renyuan Xu, Yumin Xu, and Ruixun Zhang. Diffusion factor models: Generating high-dimensional returns with factor structure. *arXiv preprint arXiv:2504.06566*, 2025. URL https://arxiv.org/abs/2504.06566.

Mincheol Cho, Yedarm Seong, and Joong-Ho Won. Pareto variational autoencoder. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=s5a8zBPFfe.

Rama Cont. Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001. doi: 10.1080/713665670.

Rama Cont, Mihai Cucuringu, Renyuan Xu, and Chao Zhang. Tail-gan: Learning to simulate tail risk scenarios. *Management Science*, 72(4):2917–2936, 2026. doi: 10.1287/mnsc.2023.00936.

Darrell Duffie and Jun Pan. An overview of value at risk. *The Journal of Derivatives*, 4(3):7–49, 1997. doi: 10.3905/jod.1997.407971.

Jianqing Fan, Han Liu, and Weichen Wang. Large covariance estimation through elliptical factor models. *The Annals of Statistics*, 46(4):1383–1414, 2018. doi: 10.1214/17-AOS1588.

Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://arxiv.org/abs/1406.2661.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Yong He, Xinbing Kong, Long Yu, and Xinsheng Zhang. Large-dimensional factor analysis without moment constraints. *Journal of Business & Economic Statistics*, 40(1):302–312, 2022. doi: 10.1080/07350015.2020.1811101.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://arxiv.org/abs/2006.11239.

Juno Kim, Jaehyuk Kwon, Mincheol Cho, Hyunjong Lee, and Joong-Ho Won. t^3-variational autoencoder: Learning heavy-tailed data with student's t and power divergence. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RzNlECeoOB.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Pavlo Krokhmal, Jonas Palmquist, and Stanislav Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of Risk*, 4(2):11–27, 2002. doi: 10.21314/JOR.2002.057.

Haoyu Liu, Tingyu Zhu, Nanshan Jia, Jinghai He, and Zeyu Zheng. Learning to simulate from heavy-tailed distribution via diffusion model. *Operations Research*, 2026. doi: 10.1287/opre.2024.1450.

James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. Don't blame the ELBO! a linear VAE perspective on posterior collapse. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://arxiv.org/abs/1911.02469.

Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, revised edition, 2015. ISBN 9780691166278.

Robert C. Merton. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1-2):125–144, 1976. doi: 10.1016/0304-405X(76)90022-2.

Sidney I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, 2007. doi: 10.1007/978-0-387-45024-7.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 1278–1286. PMLR, 2014.

Richard S. Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *International Conference on Machine Learning*, pp. 216–224, 1990. doi: 10.1016/B978-1-55860-141-3.50030-4. URL https://mlanthology.org/icml/1990/sutton1990icml-integrated/.

Hiroshi Takahashi, Tomoharu Iwata, Yuki Yamanaka, Masanori Yamada, and Satoshi Yagi. Student-t variational autoencoder for robust density estimation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 2696–2702, 2018. doi: 10.24963/ijcai.2018/374.

Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999. doi: 10.1111/1467-9868.00196.

Dacheng Xiu and Zhouyu Shen. Deep autoencoders for nonlinear factor models: Theory and applications. Research Paper 25-14, University of Chicago Booth School of Business, 2024.

A Gaussian Estimation and Tail Results

A.1 The Gaussian likelihood and ELBO

For equation (2), conditional independence across assets gives

$$\log p_\theta(y_t \mid z_t) = -\frac{1}{2}\sum_{i=1}^{N}\left[\log(2\pi) + 2\log s_i + \frac{(y_{it} - f_{\theta i}(z_t))^2}{s_i^2}\right].$$

The Gaussian variational density is positive on $\mathbb{R}^k$. Writing the marginal likelihood as an expectation and applying Jensen's inequality yields

$$\begin{aligned}\log p_\theta(y_t) &= \log \mathbb{E}_{q_\phi(z_t \mid y_t)}\left[\frac{p_\theta(y_t \mid z_t)p(z_t)}{q_\phi(z_t \mid y_t)}\right] \\ &\geq \mathbb{E}_{q_\phi(z_t \mid y_t)}[\log p_\theta(y_t \mid z_t)] - \mathrm{KL}(q_\phi(z_t \mid y_t) \,\|\, p(z_t)).\end{aligned}$$

Substitution of the likelihood and diagonal Gaussian KL gives

$$\begin{aligned}\mathcal{L}_{\mathrm{G}}(y_t) = &-\frac{1}{2}\sum_{i=1}^{N}\left[\log(2\pi) + 2\log s_i + \frac{\mathbb{E}_{q_\phi(z_t \mid y_t)}[(y_{it} - f_{\theta i}(z_t))^2]}{s_i^2}\right] \\ &-\frac{1}{2}\sum_{j=1}^{k}\left[\mu_{\phi j}(y_t)^2 + \sigma_{\phi j}(y_t)^2 - 1 - \log \sigma_{\phi j}(y_t)^2\right].\end{aligned}$$

The loss is $-T^{-1}\sum_{t=1}^{T}\mathcal{L}_{\mathrm{G}}(y_t)$. The terms involving $\log s_i$ remain part of the objective when residual scales are estimated. The identity

$$\log p_\theta(y_t) - \mathcal{L}_{\mathrm{G}}(y_t) = \mathrm{KL}(q_\phi(z_t \mid y_t) \,\|\, p_\theta(z_t \mid y_t))$$

shows that the bound is tight when the variational posterior equals the model posterior. The affine, isotropic-noise restriction is the PPCA generative model.

A.2 Proof of the Gaussian tail bound

Proof of theorem 1. Condition on the fitted parameters and fix $w \in \mathbb{R}^N$. The function $h_w(z) = w^\top f_\theta(z)$ is $L_\theta\|w\|$-Lipschitz, since

$$|h_w(z) - h_w(z')| \leq \|w\|\,\|f_\theta(z) - f_\theta(z')\| \leq L_\theta\|w\|\,\|z - z'\|.$$

Gaussian concentration states that an L-Lipschitz function h of $G \sim \mathcal{N}(0, I_k)$ satisfies

$$\mathbb{E}\exp\{\lambda(h(G) - \mathbb{E}h(G))\} \leq \exp(\lambda^2 L^2/2), \qquad \lambda \in \mathbb{R}.$$

Since $w^\top\epsilon_t$ is independent of z_t and is Gaussian with variance $w^\top Dw$, we obtain

$$\mathbb{E}e^{\lambda X_w} = \mathbb{E}e^{\lambda(h_w(z_t) - \mathbb{E}h_w(z_t))}\mathbb{E}e^{\lambda w^\top\epsilon_t} \leq \exp\left\{\frac{\lambda^2}{2}(L_\theta^2\|w\|^2 + w^\top Dw)\right\}.$$

This proves the moment-generating-function bound. Set $\tau_w^2 = L_\theta^2\|w\|^2 + w^\top Dw$. If $\tau_w^2 > 0$, Chernoff's inequality, optimized over $\lambda > 0$ for each one-sided tail, gives

$$\mathbb{P}(|X_w| > x) \leq 2\exp\left(-\frac{x^2}{2\tau_w^2}\right), \qquad x > 0.$$

Multiplication by x^α and passage to the limit proves $x^\alpha\mathbb{P}(|X_w| > x) \to 0$ for every $\alpha > 0$. If $\tau_w^2 = 0$, the moment-generating-function bound implies $X_w = 0$ almost surely, and the same conclusion follows. A target tail $\mathbb{P}(|X| > x) \sim c_0 x^{-\alpha}$ instead satisfies $x^\alpha\mathbb{P}(|X| > x) \to c_0 > 0$, proving the final assertion. □

For a feedforward decoder with weight matrices $W_1, \ldots, W_L$ and activation Lipschitz constants $\ell_1, \ldots, \ell_{L-1}$, the Lipschitz constant can be bounded as $L_\theta \leq \prod_{\ell=1}^{L}\|W_\ell\|_{\mathrm{op}}\prod_{\ell=1}^{L-1}\ell_\ell$.

B Elliptical Representations and Student-t Mixtures

B.1 Radius, direction, and Gaussian scale mixtures

An elliptical random vector in $\mathbb{R}^k$ admits a representation $m + RHU$, where m is a location vector, H determines the scatter, and U is uniform on $\mathbb{S}^{k-1}$ and independent of the nonnegative radius R (Cambanis et al., 1981). For $G \sim \mathcal{N}(0, I_k)$, the Gaussian density in polar coordinates is proportional to $r^{k-1}e^{-r^2/2}$ and does not depend on the direction. Thus $G = QU$, where $Q = \|G\| \sim \chi_k$ is independent of $U = G/\|G\|$. In dimension one, U is an independent uniform sign.

If $V \geq 0$ is independent of G, then $\sqrt{V}HG = (\sqrt{V}Q)HU$ has the radius $R = \sqrt{V}Q$, independent of U. This gives the factor and residual representations in equation (4). Gaussian scale mixtures form a subclass of elliptical distributions, with radii given by products of independent scales and chi random variables.

For the factor model, a nonsingular H can be absorbed into the unknown function f. Defining $\widetilde{F}_t = R_{f,t}U_{f,t}$ and $\widetilde{f}(x) = f(Hx)$ gives $f(R_{f,t}HU_{f,t}) = \widetilde{f}(\widetilde{F}_t)$. We therefore normalize factor scatter to I_k and retain the notation f for the resulting function. This change of coordinates leaves the return distribution unchanged.

B.2 Integrating the precisions

The Gamma density with shape $\alpha > 0$ and rate $\beta > 0$ is

$$h(\lambda; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}, \qquad \lambda > 0.$$

For $\Lambda \sim \mathrm{Gamma}(\nu/2, \nu/2)$ and $X \mid \Lambda = \lambda \sim \mathcal{N}(m, s^2/\lambda)$, integration gives

$$\begin{aligned} p(x) &= \int_0^\infty \frac{\sqrt{\lambda}}{\sqrt{2\pi}\, s} e^{-\lambda(x-m)^2/(2s^2)} \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} \lambda^{\nu/2-1} e^{-\nu\lambda/2}\, \mathrm{d}\lambda \\ &= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)\sqrt{2\pi}\, s} \frac{\Gamma((\nu+1)/2)}{[(\nu + (x-m)^2/s^2)/2]^{(\nu+1)/2}} \\ &= \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\, s} \left(1 + \frac{(x-m)^2}{\nu s^2}\right)^{-(\nu+1)/2}. \end{aligned}$$

The second line uses the Gamma integral $\int_0^\infty \lambda^{u-1}e^{-v\lambda}\, \mathrm{d}\lambda = \Gamma(u)v^{-u}$. For a k-dimensional conditional Gaussian with covariance $a^2 I_k/\lambda$, the same calculation gives

$$t_\nu^{(k)}(z; 0, a^2 I_k) = \frac{\Gamma((\nu+k)/2)}{\Gamma(\nu/2)(\nu\pi)^{k/2}a^k} \left(1 + \frac{\|z\|^2}{\nu a^2}\right)^{-(\nu+k)/2}.$$

Averaging these conditional densities over the Bernoulli state gives equations (5) and (6). The Student-t distributions thus follow from Gaussian variables divided by square roots of Gamma precisions.

B.3 Scales, variances, and tail indices

A Student-t_ν variable with scale s has variance $s^2\nu/(\nu-2)$ when $\nu > 2$. The variance ratio between the private jump and ordinary components is therefore

$$\frac{\mathrm{Var}(y_{it} \mid z_t, B_{i,t} = 1)}{\mathrm{Var}(y_{it} \mid z_t, B_{i,t} = 0)} = a_{i,1}^2 \frac{\nu_{i,1}/(\nu_{i,1}-2)}{\nu_{i,0}/(\nu_{i,0}-2)}.$$

Since the two components have the same location,

$$\mathrm{Var}(y_{it} \mid z_t) = s_i^2 \left[(1-\rho_i)\frac{\nu_{i,0}}{\nu_{i,0}-2} + \rho_i a_{i,1}^2 \frac{\nu_{i,1}}{\nu_{i,1}-2}\right].$$

Likewise,

$$\mathrm{Cov}(z_t) = \left[(1-\pi_f)\frac{\nu_{f,0}}{\nu_{f,0}-2} + \pi_f a_{f,1}^2 \frac{\nu_{f,1}}{\nu_{f,1}-2}\right] I_k.$$

The parameters in the Student-t densities specify scale or scatter. Finite factor and residual second moments, together with at most linear growth of f_θ, imply finite return second moments. For the tail calculation, the Gamma density near zero implies

$$\mathbb{P}(a/\sqrt{\Lambda} > x) = \mathbb{P}(\Lambda < a^2/x^2) \sim \frac{(\nu/2)^{\nu/2} a^\nu}{\Gamma(\nu/2+1)}\, x^{-\nu}.$$

Multiplication by the independent Gaussian radius preserves this exponent (Breiman, 1965). Equivalently, the univariate Student-t density decays as $|x|^{-\nu-1}$ and its survival probability as $x^{-\nu}$. A finite mixture with positive weights has the smallest active degree of freedom as its tail index. These are the factor and residual tail indices used in theorem 3.

C Derivation of the Variational Objective

C.1 Integrating the private variables

Conditional on z_t and the private states and precisions, the return coordinates are independent. For asset i, integration gives

$$\begin{aligned} p_\theta(y_{it} \mid z_t) = (1-\rho_i) \int_0^\infty \mathcal{N}\!\left(y_{it}; f_{\theta i}(z_t), \frac{s_i^2}{\lambda}\right) h\!\left(\lambda; \frac{\nu_{i,0}}{2}, \frac{\nu_{i,0}}{2}\right) \mathrm{d}\lambda \\ + \rho_i \int_0^\infty \mathcal{N}\!\left(y_{it}; f_{\theta i}(z_t), \frac{a_{i,1}^2 s_i^2}{\lambda}\right) h\!\left(\lambda; \frac{\nu_{i,1}}{2}, \frac{\nu_{i,1}}{2}\right) \mathrm{d}\lambda. \end{aligned}$$

The Student-t identity in section B evaluates both integrals. Independence across assets then gives $p_\theta(y_t \mid z_t) = \prod_{i=1}^N p_\theta(y_{it} \mid z_t)$, which is equation (6). The likelihood therefore accounts for every private state configuration through a product of scalar mixtures.

C.2 The common-variable ELBO

Let (g, λ, b) denote the values of $(G_t, \Lambda_{f,t}, B_{f,t})$. After integrating the private variables, the joint density is

$$p_{\theta,\psi}(y_t, g, \lambda, b) = p_\theta(y_t \mid a_{f,b}\lambda^{-1/2} g)\, \mathcal{N}(g; 0, I_k)\, \pi_f^b (1-\pi_f)^{1-b} \times h\!\left(\lambda; \frac{\nu_{f,b}}{2}, \frac{\nu_{f,b}}{2}\right).$$

Integrating g, λ and summing over b gives

$$p_{\theta,\psi}(y_t) = \int_{\mathbb{R}^k} p_\theta(y_t \mid z_t) p_\psi(z_t)\, \mathrm{d}z_t,$$

where p_ψ is the mixture prior in equation (5). For active common states, the variational density in equation (7) is positive on the support of the hierarchy. Jensen's inequality yields

$$\begin{aligned} \log p_{\theta,\psi}(y_t) = \log \mathbb{E}_{q_\phi}\left[\frac{p_{\theta,\psi}(y_t, G_t, \Lambda_{f,t}, B_{f,t})}{q_\phi(G_t, \Lambda_{f,t}, B_{f,t} \mid y_t)}\right] \geq \mathbb{E}_{q_\phi}[\log p_\theta(y_t \mid a_{f,B_{f,t}} \Lambda_{f,t}^{-1/2} G_t)] \\ - \mathrm{KL}(q_\phi(G_t, \Lambda_{f,t}, B_{f,t} \mid y_t) \,\|\, p_\psi(G_t, \Lambda_{f,t}, B_{f,t})). \end{aligned}$$

Both prior and variational densities factor into a Gaussian density, a Bernoulli mass, and a state-specific Gamma density. Taking expectations of their log ratio gives

$$\begin{aligned} &\mathrm{KL}(q_\phi(G_t, \Lambda_{f,t}, B_{f,t} \mid y_t) \,\|\, p_\psi(G_t, \Lambda_{f,t}, B_{f,t})) \\ &\quad = \mathrm{KL}(q_\phi(G_t \mid y_t) \,\|\, \mathcal{N}(0, I_k)) + \mathrm{KL}(\mathrm{Bernoulli}(r_\phi(y_t)) \,\|\, \mathrm{Bernoulli}(\pi_f)) \\ &\qquad + (1 - r_\phi(y_t))\, \mathrm{KL}(q_\phi(\Lambda_{f,t} \mid y_t, B_{f,t} = 0) \,\|\, p_\psi(\Lambda_{f,t} \mid B_{f,t} = 0)) \\ &\qquad + r_\phi(y_t)\, \mathrm{KL}(q_\phi(\Lambda_{f,t} \mid y_t, B_{f,t} = 1) \,\|\, p_\psi(\Lambda_{f,t} \mid B_{f,t} = 1)). \end{aligned}$$

The Gaussian contribution has weight $(1 - r_\phi(y_t)) + r_\phi(y_t) = 1$, because its variational density does not depend on the state. Expanding the likelihood expectation over the two states gives equation (8). Its gap to the log likelihood is

$$\log p_{\theta,\psi}(y_t) - \mathcal{L}(y_t) = \mathrm{KL}(q_\phi(G_t, \Lambda_{f,t}, B_{f,t} \mid y_t) \,\|\, p_{\theta,\psi}(G_t, \Lambda_{f,t}, B_{f,t} \mid y_t)).$$

All KL terms in the objective arise from these common latent variables. Private variables have already been integrated into the likelihood.

C.3 Analytic KL terms

For the Gaussian component,

$$\mathrm{KL}(q_\phi(G_t \mid y_t) \,\|\, \mathcal{N}(0, I_k)) = \frac{1}{2} \sum_{j=1}^{k} \left[\mu_{\phi j}(y_t)^2 + \sigma_{\phi j}(y_t)^2 - 1 - \log \sigma_{\phi j}(y_t)^2 \right].$$

For Gamma densities $q = \mathrm{Gamma}(\alpha, \beta)$ and $p = \mathrm{Gamma}(c, d)$ in shape–rate form,

$$\log \frac{q(\lambda)}{p(\lambda)} = \alpha \log \beta - c \log d - \log \Gamma(\alpha) + \log \Gamma(c) + (\alpha - c) \log \lambda - (\beta - d)\lambda.$$

The identities $\mathbb{E}_q \lambda = \alpha/\beta$ and $\mathbb{E}_q \log \lambda = \psi_0(\alpha) - \log \beta$, where ψ_0 denotes the digamma function, give

$$\mathrm{KL}(\mathrm{Gamma}(\alpha, \beta) \,\|\, \mathrm{Gamma}(c, d)) = (\alpha - c)\psi_0(\alpha) - \log \Gamma(\alpha) + \log \Gamma(c) + c \log \frac{\beta}{d} + \alpha \left(\frac{d}{\beta} - 1 \right).$$

For state b, substitute $\alpha = \alpha_{\phi,b}(y_t)$, $\beta = \beta_{\phi,b}(y_t)$, $c = d = \nu_{f,b}/2$. The Bernoulli contribution is

$$\mathrm{KL}(\mathrm{Bernoulli}(r_\phi(y_t)) \,\|\, \mathrm{Bernoulli}(\pi_f)) = (1 - r_\phi(y_t)) \log \frac{1 - r_\phi(y_t)}{1 - \pi_f} + r_\phi(y_t) \log \frac{r_\phi(y_t)}{\pi_f}.$$

C.4 Relation to the factor-space ELBO

The Gaussian scale-mixture hierarchy also induces a variational factor distribution $q_\phi(z_t \mid y_t)$. At fixed generative parameters, the transformation $(g, \lambda, b) \mapsto (z = a_{f,b} \lambda^{-1/2} g, \lambda, b)$ is invertible. Its inverse Jacobian in the continuous coordinates is $(\sqrt{\lambda}/a_{f,b})^k$. Both densities acquire this factor under the change of variables, so it cancels in their ratio and leaves the KL unchanged. Applying the KL chain rule in the transformed coordinates gives

$$\begin{aligned} \mathrm{KL}(q_\phi(G_t, \Lambda_{f,t}, B_{f,t} \mid y_t) \,\|\, p_\psi(G_t, \Lambda_{f,t}, B_{f,t})) &= \mathrm{KL}(q_\phi(z_t \mid y_t) \,\|\, p_\psi(z_t)) \\ &\quad + \mathbb{E}_{q_\phi(z_t \mid y_t)} \mathrm{KL}(q_\phi(\Lambda_{f,t}, B_{f,t} \mid z_t, y_t) \,\|\, p_\psi(\Lambda_{f,t}, B_{f,t} \mid z_t)). \end{aligned}$$

The reconstruction expectation is unchanged by the transformation. Hence, the hierarchical ELBO equals the factor-space ELBO minus the final nonnegative term. Equality holds when the conditional state and precision distributions agree for $q_\phi(z_t \mid y_t)$-almost every z_t.

D Nested Models and Reduced Objectives

In this section, we report the resulting objectives of the nested models by removing the jumps and heavy-tail components.

D.1 Removing common and private jumps

Setting $\pi_f = 0$ fixes $B_{f,t} = 0$. Finite KL then requires $r_\phi(y_t) = 0$, since the variational distribution cannot assign mass to a state excluded by the prior. The objective becomes

$$\begin{aligned} \mathcal{L}(y_t) = \mathbb{E}_{q_\phi(G_t \mid y_t) q_\phi(\Lambda_{f,t} \mid y_t, B_{f,t}=0)} &[\log p_\theta(y_t \mid \Lambda_{f,t}^{-1/2} G_t)] \\ &- \frac{1}{2} \sum_{j=1}^{k} \left[\mu_{\phi j}(y_t)^2 + \sigma_{\phi j}(y_t)^2 - 1 - \log \sigma_{\phi j}(y_t)^2 \right] \\ &- \mathrm{KL}\Big(\mathrm{Gamma}(\alpha_{\phi,0}(y_t), \beta_{\phi,0}(y_t)) \,\Big\|\, \mathrm{Gamma}\Big(\frac{\nu_{f,0}}{2}, \frac{\nu_{f,0}}{2} \Big) \Big). \end{aligned}$$

The likelihood still contains private mixtures when $\rho_i > 0$. The common jump branch, its Gamma posterior, and the Bernoulli term are omitted. This reduced expression implements the exact zero restriction without evaluating logarithms at zero.

For $\rho_i = 0$, the contribution of asset i to the decoder log likelihood is

$$\log p_\theta(y_{it} \mid z_t) = \log t_{\nu_{i,0}}(y_{it}; f_{\theta i}(z_t), s_i).$$

The common variational distribution and its KL terms are unchanged. Removing common and private jumps gives Student-t factors and Student-t residuals without mixture states.

D.2 Removing the heavy-tailed components

For $\Lambda \sim \mathrm{Gamma}(\nu/2, \nu/2)$, $\mathbb{E}\Lambda = 1$ and $\mathrm{Var}(\Lambda) = 2/\nu$, so $\Lambda \to 1$ in probability as $\nu \to \infty$. The corresponding Student-t component converges in distribution to a Gaussian. An exact Gaussian specification sets $\Lambda = 1$ and removes that variable rather than using a large finite degree of freedom. For a common Gaussian component, the reconstruction expectation no longer integrates over a precision, and its Gamma KL disappears. If there are no common jumps, the resulting objective is

$$\mathcal{L}(y_t) = \mathbb{E}_{q_\phi(G_t \mid y_t)}[\log p_\theta(y_t \mid G_t)] - \mathrm{KL}(q_\phi(G_t \mid y_t) \,\|\, \mathcal{N}(0, I_k)).$$

If all private components are also Gaussian and there are no private jumps, the likelihood is that of equation (2) and the expression is exactly equation (3).

A Gaussian ordinary component can also be paired with a Student-t jump component. In the factor prior, this gives

$$p_\psi(z_t) = (1 - \pi_f)\mathcal{N}(z_t; 0, I_k) + \pi_f t^{(k)}_{\nu_{f,1}}(z_t; 0, a^2_{f,1} I_k),$$

with objective

$$\begin{aligned}
\mathcal{L}(y_t) = (1 - r_\phi(y_t))&\mathbb{E}_{q_\phi(G_t \mid y_t)}[\log p_\theta(y_t \mid G_t)] \\
&+ r_\phi(y_t)\mathbb{E}_{q_\phi(G_t \mid y_t) q_\phi(\Lambda_{f,t} \mid y_t, B_{f,t}=1)}[\log p_\theta(y_t \mid a_{f,1}\Lambda_{f,t}^{-1/2} G_t)] \\
&- \mathrm{KL}(q_\phi(G_t \mid y_t) \,\|\, \mathcal{N}(0, I_k)) \\
&- r_\phi(y_t)\,\mathrm{KL}\Big(\mathrm{Gamma}(\alpha_{\phi,1}(y_t), \beta_{\phi,1}(y_t)) \,\Big\|\, \mathrm{Gamma}\Big(\frac{\nu_{f,1}}{2}, \frac{\nu_{f,1}}{2}\Big)\Big) \\
&- (1 - r_\phi(y_t)) \log \frac{1 - r_\phi(y_t)}{1 - \pi_f} - r_\phi(y_t) \log \frac{r_\phi(y_t)}{\pi_f}.
\end{aligned}$$

For an asset with a Gaussian ordinary component and a Student-t jump component, its likelihood is

$$p_\theta(y_{it} \mid z_t) = (1 - \rho_i)\mathcal{N}(y_{it}; f_{\theta i}(z_t), s_i^2) + \rho_i t_{\nu_{i,1}}(y_{it}; f_{\theta i}(z_t), a_{i,1} s_i).$$

These substitutions can be made separately for common factors and individual residuals.

E Training and Numerical Implementation

E.1 Reparameterized likelihood estimates

For each observation and each active common state, draw a Gaussian vector using $G_t = \mu_\phi(y_t) + \sigma_\phi(y_t) \odot \eta_t$, with $\eta_t \sim \mathcal{N}(0, I_k)$, and a precision from the corresponding Gamma posterior. The same Gaussian draw may be used in both state expectations. Each expectation is averaged over its Monte Carlo draws before multiplication by the posterior state probability. The Gaussian, Gamma, and Bernoulli KL terms are then subtracted once.

For Gamma reparameterization, let $F_{\alpha,\beta}$ be the cumulative distribution function with density $h(\,\cdot\,; \alpha, \beta)$. Differentiating $F_{\alpha,\beta}(\Lambda) = U$ at fixed $U \sim \mathrm{Unif}(0, 1)$ gives

$$\frac{\partial \Lambda}{\partial \xi} = -\frac{\partial_\xi F_{\alpha,\beta}(\Lambda)}{h(\Lambda; \alpha, \beta)}, \qquad \xi \in \{\alpha, \beta\}.$$

A differentiable Gamma sampler implements implicit reparameterization gradients without explicitly inverting the cumulative distribution (Figurnov et al., 2018). Enumerating the two common states allows ordinary differentiation through $r_\phi(y_t)$; no discrete gradient estimator is needed.

E.2 Evaluating the decoder likelihood

The univariate Student-t log density is

$$\log t_\nu(x; m, s) = \log\Gamma\Big(\frac{\nu + 1}{2}\Big) - \log\Gamma\Big(\frac{\nu}{2}\Big) - \frac{1}{2}\log(\nu\pi) - \log s - \frac{\nu + 1}{2}\log\Big(1 + \frac{(x - m)^2}{\nu s^2}\Big).$$

Log-Gamma and `log1p` functions evaluate the normalizer and final term. For a private mixture, use

$$\begin{aligned}
\log p_\theta(y_{it} \mid z_t) = \mathrm{logaddexp}\big(&\log(1 - \rho_i) + \log t_{\nu_{i,0}}(y_{it}; f_{\theta i}(z_t), s_i), \\
&\log \rho_i + \log t_{\nu_{i,1}}(y_{it}; f_{\theta i}(z_t), a_{i,1} s_i)\big).
\end{aligned}$$

Summing these values across assets gives the joint log likelihood. This product-of-mixtures calculation is linear in N; it does not enumerate the 2^N private state vectors. Exact zero weights and Gaussian components use the reduced expressions in section D.

E.3 Parameter constraints and moment conditions

The decoder mean is a feedforward network with real-valued outputs, and s_i are positive asset-specific parameters. The encoder outputs k Gaussian means, k positive Gaussian scales, a common-state probability when jumps are active, and a shape and rate for each active Gamma component. Positive outputs can be parameterized with softplus and a fixed positive floor. The posterior probability $r_\phi(y_t)$ ranges over $(0, 1)$ even when the prior jump probability is constrained to $(0, 1/2)$.

Hyperparameter selection. We tune the factor dimension, mixture probabilities (π_f, ρ_i), jump-scale multipliers $(a_{f,1}, a_{i,1})$, and degrees of freedom $(\nu_{f,b}, \nu_{i,b})$ through a restricted grid search. Across the two markets, candidate values collectively span $k \in [4, 10]$, $\pi_f \in [0.001, 0.0075]$, $\rho_i \in [0.001, 0.004]$, $a_{f,1}, a_{i,1} \in [2, 3]$, $\nu_{f,0} \in [6, 20]$, $\nu_{i,0} \in [4, 20]$, $\nu_{f,1} \in [4, 6]$, and $\nu_{i,1} \in [4, 5]$. These ranges summarize the discrete candidate configurations considered across the two markets; all parameter combinations were evaluated. Candidate configurations are assessed at the validation window stages to examine the consistency of tail accuracy, distributional fit, and portfolio performance before extending the evaluation to the full rolling horizon. The selected configuration is fixed across observations and rolling fits within each dataset. Private radial hyperparameters are shared across assets, with $a_{f,0} = a_{i,0} = 1$ and the component ordering in section 2.3. Within each fitting window, the encoder and decoder parameters and residual scales s_i are learned, while the encoder produces observation-specific posterior parameters.

Moment conditions depend on the specification. For a Gamma posterior, $\mathbb{E}[\Lambda^{-u}]$ is finite when its shape exceeds u. Thus active posterior shapes above $1/2$ give finite factor means, and shapes above one give finite second moments. With an at-most-linearly growing decoder, shapes above one are sufficient for finite expected squared reconstruction error whenever a residual likelihood is purely Gaussian. Student-t negative log likelihoods grow only logarithmically in the residual magnitude and do not impose this same second-moment condition.

The nonlinear decoder uses SiLU hidden activations and an unconstrained linear output layer. Unlike bounded activations such as tanh or sigmoid, SiLU is unbounded and asymptotically linear on the positive axis, allowing heavy-tailed latent factors to produce unbounded common returns. Preservation of the latent tail index additionally requires a non-degenerate asymptotic decoder response on directions carrying positive tail mass.

E.4 Objective normalization and sampling

Within each observation, all N asset log likelihoods are summed and one complete common-factor KL is subtracted. The resulting ELBO is then averaged over the minibatch. Averaging only the likelihood over assets would change its weight relative to the KL. Increasing the number of Monte Carlo draws changes the average used to estimate the likelihood expectation. If KL warm-up is used, the final fitting objective uses unit weight.The experiment-specific settings are collected in section H.

F Posterior Summaries for Financial Interpretation

F.1 Common factors and scales

For an observed return vector y_t, the variational probability $r_\phi(y_t)$ summarizes the common jump state. The expected log variance multiplier is

$$\begin{aligned}\mathbb{E}_{q_\phi}[\log V_{f,t} \mid y_t] = (1 - r_\phi(y_t))\,[\log \beta_{\phi,0}(y_t) - \psi_0(\alpha_{\phi,0}(y_t))] \\ + r_\phi(y_t)\,[2\log a_{f,1} + \log \beta_{\phi,1}(y_t) - \psi_0(\alpha_{\phi,1}(y_t))]\,.\end{aligned}$$

This exists for all positive Gamma shapes and rates. If the active posterior shapes exceed one, the mean multiplier is

$$\mathbb{E}_{q_\phi}[V_{f,t} \mid y_t] = (1 - r_\phi(y_t))\frac{\beta_{\phi,0}(y_t)}{\alpha_{\phi,0}(y_t) - 1} + r_\phi(y_t) a_{f,1}^2 \frac{\beta_{\phi,1}(y_t)}{\alpha_{\phi,1}(y_t) - 1}.$$

The factor moments also depend on the Gaussian posterior. When active shapes exceed $1/2$, independence of G_t and $\Lambda_{f,t}$ within each state gives

$$\mathbb{E}_{q_\phi}[z_t \mid y_t] = \mu_\phi(y_t)\Bigg[(1 - r_\phi(y_t))\sqrt{\beta_{\phi,0}(y_t)}\frac{\Gamma(\alpha_{\phi,0}(y_t) - 1/2)}{\Gamma(\alpha_{\phi,0}(y_t))} + r_\phi(y_t)a_{f,1}\sqrt{\beta_{\phi,1}(y_t)}\frac{\Gamma(\alpha_{\phi,1}(y_t) - 1/2)}{\Gamma(\alpha_{\phi,1}(y_t))}\Bigg].$$

With shapes above one,

$$\mathbb{E}_{q_\phi}[z_t z_t^\top \mid y_t] = \mathbb{E}_{q_\phi}[V_{f,t} \mid y_t]\ \left(\operatorname{diag}\{\sigma_\phi^2(y_t)\} + \mu_\phi(y_t)\mu_\phi(y_t)^\top\right).$$

Subtracting the outer product of the posterior mean gives $\operatorname{Cov}_{q_\phi}(z_t \mid y_t)$. These expressions distinguish factor uncertainty from uncertainty in its multiplier. Posterior draws and quantiles can be used when moments do not exist.

F.2 Private jump probabilities

Conditional on z_t, the private jump responsibility is

$$\mathbb{P}(B_{i,t} = 1 \mid y_{it}, z_t) = \frac{\rho_i t_{\nu_{i,1}}(y_{it}; f_{\theta i}(z_t), a_{i,1}s_i)}{(1-\rho_i)t_{\nu_{i,0}}(y_{it}; f_{\theta i}(z_t), s_i) + \rho_i t_{\nu_{i,1}}(y_{it}; f_{\theta i}(z_t), a_{i,1}s_i)}. \tag{10}$$

Averaging equation (10) over the induced $q_\phi(z_t \mid y_t)$ gives a variational estimate of the private-jump probability. The conditional responsibility is exact, while the factor posterior and its Monte Carlo average introduce approximation.

G Extreme Directions and the Common Loading Subspace

G.1 Assumptions and the full tail result

All limits below hold for fixed N and k, with the Euclidean norm. For one period, suppress the time index in equation (9) and write $y - c = C + e$, where $C = \sqrt{V_f}AG$ and $e_i = s_i\sqrt{V_i}E_i$. Here $A \in \mathbb{R}^{N\times k}$ is fixed, $G \sim \mathcal{N}(0, I_k)$, and $E_i \sim \mathcal{N}(0,1)$. The nonnegative multipliers $V_f, V_1, \dots, V_N$, the vector G, and the scalar variables $E_1, \dots, E_N$ are mutually independent.

A random vector $X \in \mathbb{R}^N$ is multivariate regularly varying with index $\alpha > 0$ and spectral probability measure H if, for every $r > 0$ and every continuity set $B \subset \mathbb{S}^{N-1}$ of H,

$$\frac{\mathbb{P}(\|X\| > rx,\ X/\|X\| \in B)}{\mathbb{P}(\|X\| > x)} \longrightarrow r^{-\alpha}H(B), \qquad x \to \infty.$$

Thus the radial tail describes the frequency of extremes, while H describes their directions. We use H_y for the spectral measure of the centered return $y - c$.

Theorem 3 (Common and private extreme directions)**.** *Under the independence assumptions above, let $A \neq 0$ and $s_i > 0$. Suppose*

$$\mathbb{P}(\sqrt{V_f} > x) \sim c_f x^{-\alpha_f}, \qquad \mathbb{P}(\sqrt{V_i} > x) \sim c_i x^{-\alpha_i}, \qquad x \to \infty,$$

where all constants and exponents are positive. Define $\alpha_e = \min_i \alpha_i$ and $I_e = \{i : \alpha_i = \alpha_e\}$. Then C, e, and $y - c$ are multivariate regularly varying with indices α_f, α_e, and $\alpha = \min(\alpha_f, \alpha_e)$, respectively. Let H_f and H_e be the spectral measures of C and e. The return spectral measure is

$$H_y = \omega H_f + (1-\omega)H_e, \qquad \omega = \begin{cases} 1, & \alpha_f < \alpha_e, \\ d_f/(d_f + d_e), & \alpha_f = \alpha_e, \\ 0, & \alpha_f > \alpha_e, \end{cases}$$

where

$$d_f = c_f\mathbb{E}\|AG\|^{\alpha_f}, \qquad d_i = c_i s_i^{\alpha_i}\mathbb{E}|E_i|^{\alpha_i}, \qquad d_e = \sum_{i\in I_e} d_i.$$

The measures H_f and H_e are given explicitly in equations (11) and (12); their contributions to H_y are summarized in table 3.

Table 3: Extreme directions under theorem 3. A smaller tail index means a heavier tail. The common loading subspace is $\mathrm{range}(A)$; v_i denotes the ith coordinate unit vector. In the equal-index case, $0 < \omega < 1$.

Tail comparison	Limiting measure H_y	Support of the extreme directions
$\alpha_f < \alpha_e$	H_f	$\mathrm{range}(A) \cap \mathbb{S}^{N-1}$
$\alpha_e < \alpha_f$	H_e	$\{\pm v_i : i \in I_e\}$
$\alpha_f = \alpha_e$	$\omega H_f + (1-\omega) H_e$	Union of the common and private supports

G.2 Common extremes

Let $W = AG$. Since $A \neq 0$, $W \neq 0$ almost surely, and Gaussian moments of every positive order are finite. Conditioning on W gives

$$\mathbb{P}(\|C\| > x,\ C/\|C\| \in B) = \mathbb{E}\left[\mathbf{1}_{\{W/\|W\| \in B\}} \mathbb{P}\Big(\sqrt{V_f} > x/\|W\| \;\Big|\; W\Big)\right].$$

The pure power tail assumption implies a finite constant K such that $\mathbb{P}(\sqrt{V_f} > u) \leq K u^{-\alpha_f}$ for every $u > 0$. After multiplication by x^{α_f}, the integrand is therefore bounded by $K\|W\|^{\alpha_f}$. Dominated convergence yields

$$\mathbb{P}(\|C\| > x) \sim d_f x^{-\alpha_f}, \quad H_f(B) = \frac{\mathbb{E}[\|AG\|^{\alpha_f} \mathbf{1}_{\{AG/\|AG\| \in B\}}]}{\mathbb{E}\|AG\|^{\alpha_f}}. \tag{11}$$

Replacing x by rx gives the full regular-variation limit. Equivalently, writing $G = \chi_k U_f$ and cancelling the independent Gaussian radial moment gives

$$H_f(B) = \frac{\mathbb{E}[\|AU_f\|^{\alpha_f} \mathbf{1}_{\{AU_f/\|AU_f\| \in B\}}]}{\mathbb{E}\|AU_f\|^{\alpha_f}}.$$

Thus directions amplified more by A receive more tail weight. The Gaussian vector W has a positive density on $\mathrm{range}(A)$ with respect to Lebesgue measure on that subspace. Its radial weight is positive away from zero, so $\mathrm{supp}(H_f) = \mathrm{range}(A) \cap \mathbb{S}^{N-1}$, including when A has deficient column rank.

G.3 Private extremes

The same conditioning argument gives, for each coordinate,

$$\mathbb{P}(|e_i| > x) \sim d_i x^{-\alpha_i}.$$

Coordinates outside I_e have negligible tails on the $x^{-\alpha_e}$ scale. Independence implies, for distinct i, j and every fixed $\eta > 0$,

$$\mathbb{P}(|e_i| > \eta x,\ |e_j| > \eta x) = O(x^{-\alpha_i - \alpha_j}) = o(x^{-\alpha_e}).$$

Consequently, at leading order, only one private coordinate is of order x. Within that event, the remaining coordinates change the scaled vector by at most $\sqrt{N}\eta$ when all are bounded by ηx. Summing the single-coordinate tail contributions and then letting $\eta \downarrow 0$ gives

$$\mathbb{P}(\|e\| > x) \sim d_e x^{-\alpha_e}, \quad H_e = \sum_{i \in I_e} \frac{d_i}{d_e} \frac{\delta_{v_i} + \delta_{-v_i}}{2}. \tag{12}$$

Here δ_v is a unit point mass at direction v. Symmetry gives equal positive and negative weights. Since the Gaussian absolute moment is common across $i \in I_e$, the normalized asset weight also equals $c_i s_i^{\alpha_e} / \sum_{j \in I_e} c_j s_j^{\alpha_e}$.

G.4 Adding the common and private tail measures

The vectors C and e are independent. If $\alpha_f < \alpha_e$, then for every fixed $\eta > 0$,

$$\frac{\mathbb{P}(\|e\| > \eta x)}{\mathbb{P}(\|C\| > x)} \longrightarrow 0.$$

This perturbation does not change the common tail measure. To see this, take a continuous function h with compact support bounded away from zero. On $\{\|e\| \le \eta x\}$, the arguments of $h(x^{-1}(C+e))$ and $h(x^{-1}C)$ differ by at most η. Their difference is bounded by the modulus of continuity of h and can be nonzero only if $\|C\|$ exceeds a positive multiple of x, provided η is sufficiently small. The probability of that event is $O(x^{-\alpha_f})$. The complement $\{\|e\| > \eta x\}$ has probability $o(x^{-\alpha_f})$. First let $x \to \infty$, then $\eta \downarrow 0$. The two scaled expectations have the same limit, proving $H_y = H_f$ and tail index α_f. The case $\alpha_e < \alpha_f$ is symmetric.

If $\alpha_f = \alpha_e = \alpha$, independence gives

$$\mathbb{P}(\|C\| > \eta x,\ \|e\| > \eta x) = O(x^{-2\alpha}) = o(x^{-\alpha}).$$

Consider $h(x^{-1}(C+e)) - h(x^{-1}C) - h(x^{-1}e)$. When both norms exceed ηx, boundedness of h and the preceding probability bound make its contribution negligible. When neither exceeds ηx, the expression vanishes for sufficiently small η, because the support of h is away from zero. When exactly one exceeds ηx, uniform continuity gives the same vanishing error bound as above. Thus the limiting tail measure is the sum of the common and private tail measures. The bound

$$\mathbb{P}(\|C+e\| > Rx) \le \mathbb{P}(\|C\| > Rx/2) + \mathbb{P}(\|e\| > Rx/2)$$

controls mass beyond large R, so these limits also give the radial tail constants and normalized spectral measures. In particular,

$$\mathbb{P}(\|y-c\| > x) \sim (d_f + d_e)x^{-\alpha}, \quad H_y = \frac{d_f}{d_f+d_e}H_f + \frac{d_e}{d_f+d_e}H_e.$$

All constants are positive, and therefore

$$\omega = \frac{c_f \mathbb{E}\|AG\|^\alpha}{c_f \mathbb{E}\|AG\|^\alpha + \mathbb{E}|E_1|^\alpha \sum_{i\in I_e} c_i s_i^\alpha} \in (0,1).$$

Together with equations (11) and (12), the three cases prove theorem 3. The product and summation arguments are specializations of standard regular-variation results (Breiman, 1965; Basrak et al., 2002; Resnick, 2007).

G.5 Proof of the fixed common subspace

We now prove the matrix statements in theorem 2. Let $\mathcal{S} = \mathrm{range}(A)$ and $r = \mathrm{rank}(A)$. Full column rank is not required. From equation (11),

$$M_f := \int uu^\top H_f(\mathrm{d}u) = \frac{\mathbb{E}[\|W\|^{\alpha_f - 2} WW^\top]}{\mathbb{E}\|W\|^{\alpha_f}}, \qquad W = AG.$$

The numerator is finite even when $\alpha_f < 2$: the norm of its integrand is $\|W\|^{\alpha_f}$, which has finite, positive expectation. For every $a \in \mathbb{R}^N$,

$$a^\top M_f a = \frac{\mathbb{E}[\|W\|^{\alpha_f - 2}(a^\top W)^2]}{\mathbb{E}\|W\|^{\alpha_f}}.$$

This expression is zero if $A^\top a = 0$. Otherwise $a^\top W$ is a nondegenerate Gaussian, and the integrand is strictly positive almost surely. Consequently,

$$\ker(M_f) = \ker(A^\top) = \mathcal{S}^\perp, \qquad \mathrm{range}(M_f) = \mathcal{S}, \qquad \mathrm{rank}(M_f) = r.$$

The private spectral measure in equation (12) gives the diagonal matrix

$$D_e := \int uu^\top H_e(\mathrm{d}u) = \sum_{i\in I_e} \frac{d_i}{d_e} v_i v_i^\top.$$

Integrating $uu^\top$ against the spectral mixture in theorem 3 proves

$$M_{\mathrm{tail}} := \int uu^\top H_y(\mathrm{d}u) = \omega M_f + (1-\omega) D_e. \tag{13}$$

When the second moments are finite, the independence assumptions and the centered Gaussian variables also give

$$\mathrm{Cov}(y) = \tau_f AA^\top + D, \qquad \tau_f = \mathbb{E}[V_f] > 0, \qquad D = \mathrm{diag}\{s_i^2 \mathbb{E}[V_i]\}_{i=1}^N.$$

Since $\mathrm{range}(AA^\top) = \mathcal{S}$, the common covariance component has a range $\mathcal{S}$. If $\omega > 0$, the same holds for the common tail component ωM_f, proving theorem 2. If $\omega = 0$, that component vanishes from the limiting tail measure, even though the common loading subspace can be nontrivial.

H Experimental Details and Additional Results

H.1 Synthetic design

Each of three settings contains $T = 6{,}000$ independent observations of $N = 60$ coordinates with $k = 5$ common factors:

$$z_t = \sqrt{V_{f,t}} G_t, \quad G_t \sim \mathcal{N}(0, I_k), \quad E_{it} \sim \mathcal{N}(0,1),$$
$$y_{it} = \mu_i + d_i \left\{ [g_s(z_t)]_i + s_i \sqrt{V_{i,t}} E_{it} \right\}.$$

The scale multipliers follow the two-component laws in section 2.3, with parameters in table 4. Private mixture parameters are shared across assets. Gaussian variables and scale multipliers are independent across observations and sources, and private primitives are independent across coordinates.

Table 4: Scale-mixture parameters shared by all synthetic settings. Ordinary-state relative scales equal one; private parameters are shared across coordinates.

Parameter	Common factors	Private residuals
Ordinary degrees of freedom	8	7
Jump degrees of freedom	4.5	4.5
Jump probability	0.01	0.003
Jump relative scale	2.5	3

All settings share these scale laws and differ only in their observation maps. With width $m = 20$ and $h(z) = \tanh(Az) \in \mathbb{R}^m$,

$$g_1(z) = Wz,$$
$$g_2(z) = Wz + 0.35Bh(z),$$
$$g_3(z) = Wz + 0.35Bh(z) + 0.55D\tanh(Ch(z)) + 0.20B[h(z) \odot h(z)].$$

The matrices have dimensions $W \in \mathbb{R}^{N\times k}$, $A \in \mathbb{R}^{m\times k}$, $B, D \in \mathbb{R}^{N\times m}$, and $C \in \mathbb{R}^{m\times m}$. Entries are independent centered Gaussians with variances $0.55^2/k$, $0.75^2/k$, $0.45^2/m$, and $1/m$ for W, A, (B, D), and C, respectively. Matrices are drawn once and fixed. We draw $\mu_i \sim \mathrm{Unif}(0.00005, 0.00035)$ and $s_i \sim \mathrm{Unif}(0.35, 0.65)$. An independent 10,000-observation pilot sets

$$d_i = \frac{0.015}{\widehat{\mathrm{sd}}_{\mathrm{pilot}}\left([g_s(z)]_i + s_i\sqrt{V_i}E_i\right)}$$

before fitting and evaluation observations are generated. S1 is linear; S2 adds a bounded nonlinear response; S3 adds nonlinear depth and an even response component. The additional terms are bounded, so all three maps retain the asymptotic linear component Wz. This design therefore assesses nonlinear fit at finite magnitudes.

Implementation details. For G-VAE and the E-VAE variants, the following training procedure is used in both the synthetic and real-data experiments. SiLU activations are used in the VAE hidden layers. Parameters are optimized using AdamW with a learning rate of 0.002 and a batch size of 128. Weight decay of 10^{-5} is applied to network parameters but not to scale, probability, or Gamma parameter groups. Training is limited to 80 epochs, with a 20-epoch KL warm-up and an early-stopping patience of 15 epochs. The checkpoint with the lowest validation negative ELBO at unit KL weight is retained. Training and validation expectations are estimated using 4 and 32 Monte Carlo draws, respectively, and gradient norms are clipped at 10. Linear layers are initialized using the PyTorch defaults, while Gamma and jump heads are initialized to their prior settings. Log scales are constrained to $[-6, 2]$, Gamma shapes to values above one, and Gamma rates to values of at least 10^{-3}.

Tail-GAN is trained for 3,000 epochs with five updates per epoch, a batch size of 256, and generator and critic learning rates of 10^{-4} and 10^{-5}, respectively. Its tail level is fixed at 5%, using 32 random portfolios and 10 fitted eigenportfolios; the final checkpoint is retained. Tail-Diffusion is trained for 300 epochs with a learning rate decreasing from 0.01 to 10^{-5}. Sampling is performed over 20 noise

levels with 20 updates per level, using Student-t noise with 3.5 degrees of freedom. Both baselines are adapted to daily return vectors and generate 4,000 scenarios per evaluation window.

In the synthetic experiments, t^3-VAE and Pareto-VAE use the same training, validation, and test observations as the other models in each rolling window. Both use five latent factors and three hidden layers of 64 units with SiLU activations, matching the architecture of G-VAE and the E-VAE variants. We use tabular adaptations of the publicly released implementations. For t^3-VAE, we use a Student-t specification with $\nu = 5$, following the default in its publicly released synthetic-experiment scripts. This value lies between the jump-component degrees of freedom (4.5) and those of the ordinary factor and residual components (8 and 7, respectively) in our data-generating process. Pareto-VAE uses a symmetric multivariate Pareto specification with base tail parameter $\nu = 3.1$ and an ℓ_1 reconstruction term, with corrections to align its sampler and loss with the paper. Both models are trained using their model-specific power-divergence objectives with AdamW, a learning rate of 0.002, and a batch size of 128. The epoch limits are 100 and 200, with early-stopping patience of 15 and 100, respectively. Checkpoints are selected using each model's validation objective. All nine models generate 4,000 observations per window and are evaluated using the same sampling protocol and performance measures.

H.2 Evaluation samples and distributional metrics

The same evaluation protocol is used for synthetic and real data. For each window, generated samples are drawn with replacement from 4,000 scenarios, using 25 samples per held-out observation for quantile evaluation and four for distributional metrics. Windows are therefore weighted by their held-out sample sizes. Empirical quantiles are computed by linear interpolation.

Standardization. For distributional metrics, held-out and generated returns are standardized using the coordinate means and population standard deviations from the first fitting window:

$$x = D_{\mathrm{fit}}^{-1}(y - \mu_{\mathrm{fit}}), \qquad \widetilde{x} = D_{\mathrm{fit}}^{-1}(\widetilde{y} - \mu_{\mathrm{fit}}),$$

with standard deviations floored at 10^{-8}. The norm CDFs are computed from $\|x\|_2$ and $\|\widetilde{x}\|_2$.

Fréchet discrepancy. For empirical means $\bar{x}, \bar{\widetilde{x}}$ and sample covariances $S_x, S_{\widetilde{x}}$, the normalized discrepancy is

$$\frac{\mathrm{FD}}{N} = \frac{1}{N}\left[\|\bar{x} - \bar{\widetilde{x}}\|_2^2 + \mathrm{tr}\left\{S_x + S_{\widetilde{x}} - 2\left(S_x^{1/2} S_{\widetilde{x}} S_x^{1/2}\right)^{1/2}\right\}\right].$$

Maximum mean discrepancy. A Gaussian kernel mixture is used:

$$k(u, v) = \frac{1}{3}\sum_{c \in \{0.25, 1, 4\}} \exp\left(-\frac{\|u - v\|_2^2}{2c\ell^2}\right),$$

where ℓ^2 is the median pairwise squared distance among at most 512 standardized fitting observations, floored at 10^{-8} and shared across models. For two samples of size r,

$$\widehat{\mathrm{MMD}}_u^2 = \frac{\sum_{a \neq b} k(x_a, x_b) + \sum_{a \neq b} k(\widetilde{x}_a, \widetilde{x}_b)}{r(r-1)} - \frac{2}{r^2}\sum_{a,b} k(x_a, \widetilde{x}_b).$$

Negative estimates are retained; reported values are scaled by 10^3.

Energy discrepancy. The empirical energy discrepancy is computed as

$$\widehat{\mathrm{Energy}}_v = \frac{1}{r^2}\sum_{a,b}\left(2\|x_a - \widetilde{x}_b\|_2 - \|x_a - x_b\|_2 - \|\widetilde{x}_a - \widetilde{x}_b\|_2\right).$$

Sliced Wasserstein distance. Using $L = 128$ independent directions drawn uniformly from the unit sphere, the distance is estimated by

$$\widehat{\mathrm{SW}}_1 = \frac{1}{Lr}\sum_{\ell=1}^{L}\sum_{a=1}^{r}\left|x_{(a)}^{(\ell)} - \widetilde{x}_{(a)}^{(\ell)}\right|,$$

where $x_{(a)}^{(\ell)}$ and $\widetilde{x}_{(a)}^{(\ell)}$ are the ordered projections onto direction ℓ.

FD is computed using the full standardized samples. MMD, Energy, and SW1 are averaged over three subsampling replicates, each using $r = \min(1024, n, m)$ vectors drawn without replacement from each pool, where n and m are the held-out and generated pool sizes. The same subsampling protocol is applied to all models.

H.3 Market data and fitting settings

Observations are grouped by quarter, the out-of-sample test periods span February 1, 2005–December 31, 2020 in the U.S. market (64 windows; 4,008 trading days) and August 1, 2005–December 31, 2020 in the Chinese market (62 windows; 3,752 trading days).

Daily excess returns are expressed in decimal units. One U.S. company with incomplete observations is excluded, leaving 158 assets. Within each window, returns are standardized using means and sample standard deviations estimated from the training subset only. The fixed normalization used for distributional evaluation is described in section H.2.

For both markets, eight latent factors and three hidden layers of width 64 with SiLU activations are used in the encoder and decoder. The remaining training settings are shared with the synthetic experiments.

In both the common and private components, ordinary-state degrees of freedom are fixed at 20, jump-state degrees of freedom at 5, and jump scale multipliers at 2. The common and private prior jump probabilities are set to 0.005 and 0.0025 for the U.S. market, and 0.0075 and 0.0015 for the Chinese market, respectively. These values apply when the corresponding jump component is included; otherwise, its probability is set to zero. All distributional hyperparameters are held fixed during training, while the encoder and decoder parameters and asset-specific residual scales are jointly optimized.

H.4 Portfolio construction

Portfolio construction. For return vector y, portfolio loss is $-w^\top y$. The feasible set is

$$\mathcal{W} = \{w : \mathbf{1}^\top w = 1,\ 0 \le w_i \le 0.05\}.$$

For each model and evaluation window, 1,000 scenarios are selected by taking every fourth vector from the generated pool of 4,000. The minimum estimated tail risk is first computed:

$$v_* = \min_{w \in \mathcal{W}} \widehat{\mathrm{CVaR}}_{0.95}(w).$$

The target portfolio is then obtained by solving

$$\begin{aligned} w^* \in \arg\max_{w \in \mathcal{W}} \quad & \widehat{\mu}^\top w - \eta \|w - w_{\mathrm{prev}}\|_1 \\ \text{subject to} \quad & \widehat{\mathrm{CVaR}}_{0.95}(w) \le 1.1 \max(v_*, 10^{-8}), \end{aligned}$$

where $\widehat{\mu}$ is the annualized scenario mean (252 times the daily mean), $\eta = 0.005$, and CVaR is measured in daily loss units. Thus, expected return is maximized under a tail-risk budget with a turnover penalty. The problems are solved in CVXPY using CLARABEL, with SCS as a fallback, warm starts, and default solver tolerances.

Return accounting and performance metrics. Initial weights are equal, and quarterly target weights are restored daily. Let w_t^{pre} and w_t denote pre-trade and target weights. Daily turnover is $\tau_t = \|w_t - w_t^{\mathrm{pre}}\|_1$, and net excess return is $r_t^{\mathrm{net}} = w_t^\top y_t - 0.002\tau_t$. The next day's pre-trade weights are

$$w_{i,t+1}^{\mathrm{pre}} = \frac{w_{it}(1 + y_{it})}{\sum_j w_{jt}(1 + y_{jt})}.$$

Weight drift is computed using excess returns, with transaction costs deducted only from net returns.

For n evaluation days, let $\bar{r}$ and s_r denote the mean and sample standard deviation of daily net excess returns. Annualized mean return and Sharpe ratio are $\mathrm{Mean/yr} = 252\bar{r}$ and $\mathrm{SR} = \sqrt{252}\,\bar{r}/s_r$. With

$W_t = \prod_{s=1}^{t}(1 + r_s^{\text{net}})$ and $W_0 = 1$, cumulative return is $\text{CR} = W_n - 1$ and maximum drawdown is $\text{MDD} = \max_{1 \le t \le n}\{1 - W_t / \max_{0 \le s \le t} W_s\}$. ES95 averages daily losses $-r_t^{\text{net}}$ at or above their empirical 95% quantile, including ties. Annualized trading cost is $\text{Cost/yr} = 252\,\overline{0.002\tau_t}$. All measures except SR are reported as percentages. CR and MDD use compounded net excess returns rather than total-return wealth.

H.5 Historical posterior scale and factor summaries

Both posterior moments average over the ordinary and jump states. Models are fitted before each evaluation window; all three series are computed for each observation and then smoothed with a 21-observation trailing mean.

RMS values are computed before smoothing; trailing averages continue across evaluation windows and use the available history for the first 20 observations.

The shaded intervals in figure 3 mark the global financial crisis (December 2007-June 2009), the Chinese equity selloff (June 8-August 26, 2015), early-2016 Chinese market turbulence (January-February 2016), and COVID-19 (February-April 2020 in the U.S.; January 20-March 31, 2020 in China). China-specific episodes are marked only in the Chinese panel.

H.6 Computational measurements

E-VAE and Tail-GAN are evaluated on different systems. E-VAE is run on a 10-core Apple M5 CPU with macOS 26.6.2 (ARM64) and PyTorch 2.13.0, using five concurrent model workers and one PyTorch thread per fit. Tail-GAN is run on an NVIDIA GeForce RTX 5060 Ti GPU with 16 GB of memory, using PyTorch 2.11.0, CUDA 12.8, and NumPy 2.2.6.

For E-VAE-CP, the mean end-to-end runtime is 30.81 s per window across 64 windows, and the generation of 4,000 return vectors takes 0.033 s. For Tail-GAN, training takes 1,640.76 s per window, including validation and checkpoint saving, and the generation of 4,000 return vectors takes 0.02 s. These are implementation-specific runtimes measured on the respective systems described above.

H.7 Body-tail geometry and decoder linearity

Body and tail subspaces. Returns are standardized using the fixed normalization in section H.2. For each nonzero standardized vector x_t, define $u_t = x_t / \|x_t\|_2$. Within each evaluation window of size n_w, the $\lfloor 0.8 n_w \rfloor$ smallest and $\lceil 0.1 n_w \rceil$ largest norms are selected as body and tail observations, respectively, with ties resolved chronologically. Selected directions are pooled across windows without further centering or whitening. For either pooled set $\mathcal{S}$, the angular second-moment matrix is $M_{\mathcal{S}} = |\mathcal{S}|^{-1} \sum_{t \in \mathcal{S}} u_t u_t^\top$. Its cumulative eigenvalues measure angular mass captured by leading subspaces, while $\sum_i u_{it}^4$ measures coordinate concentration.

Let Q_b, Q_t contain the leading $k = 8$ eigenvectors of the body and tail matrices. Principal angles are defined by $\theta_j = \arccos(s_j)$, where s_j are the descending singular values of $Q_b^\top Q_t$. The aggregate subspace gap is $\{1 - \|Q_b^\top Q_t\|_F^2 / k\}^{1/2}$.

Affine versus nonlinear decoding. Figure 4 reports the principal-angle sines $\sin(\theta_j)$ for observed U.S. and Chinese returns alongside the decoder comparison, using the same dates across panels within each market. The figure uses the full out-of-sample test periods for both the U.S. and Chinese stock markets. Tail observations are selected by return magnitude and include both positive and negative extremes.

For the decoder comparison, the nonlinear mean function is replaced by $f_\theta(z) = c + Az$, while the encoder architecture, factor dimension, radial settings, data splits, standardization, optimization, training budget, and window-specific seed are held fixed. Each model is fitted separately and evaluated using the same sampling protocol and metrics. Full-period results are reported in table 5. Nonlinear decoding reduces all four lower-tail QMAEs and all four joint distributional discrepancies in both markets.

Table 5: Paired E-VAE-CP decoder ablation over the full evaluation periods. Lower is better. Qα denotes QMAE at the $\alpha\%$ quantile, in return pp; MMD^2 is multiplied by 10^3.

Market	Decoder	Q1	Q2.5	Q5	Q7.5	FD/N	MMD^2	Energy	SW1
U.S.	nonlinear	1.295	0.448	0.161	0.164	0.076	1.148	0.055	0.094
U.S.	affine	1.862	0.916	0.477	0.295	0.157	3.392	0.113	0.127
Chinese	nonlinear	1.995	0.595	0.214	0.372	0.163	2.616	0.109	0.149
Chinese	affine	3.265	1.613	0.803	0.459	0.333	4.529	0.159	0.193

I USE OF LANGUAGE MODELS

Language-model assistance was used in checking derivations, code prototyping, and manuscript editing. Responsibility for scientific decisions and verification of the reported results remains with the authors.