Variational Bayes for Limit Order Book Estimation: A Probabilistic Approach to High-Frequency Trading Dynamics

AMS 522 Project report

Matvei Lukianov
Quantitative Finance
Stony Brook University
matvei.lukianov@stonybrook.edu

May 4, 2025

Contents

1 Introduction

High-frequency trading (HFT) has reshaped modern financial markets by enabling the execution of massive trade volumes at sub-second latencies. Paramount to the success of HFT strategies is an accurate representation of the *fair price*, an unobserved latent variable that reflects the market's consensus value around which buys and sells cluster. I treat this fair price as a latent variable, denoted Z, which drives both the ask and bid sides of the limit order book (LOB).

The LOB records outstanding buy (bid) and sell (ask) orders at discrete price levels, revealing granular information on liquidity, depth, and short-term price dynamics. Traditional approaches—such as maximum-likelihood estimation or discrete-time Markov models—often sacrifice either computational tractability or the ability to quantify uncertainty when applied to rapidly evolving LOB data. Bayesian methods naturally address uncertainty and offer principled posterior inference, but classic sampling approaches like Markov Chain Monte Carlo (MCMC) can become a bottleneck: MCMC typically yields highly accurate posterior samples of the fair price Z yet can be too slow for real-time HFT applications.

Variational Bayes (VB) emerges as a scalable alternative, approximating intractable posteriors with simpler distributions that update via efficient optimization. In this study, I develop a VB framework to infer the latent fair price shocks on both ask and bid sides, and I compare its performance against MCMC—the benchmark for inference accuracy. I apply both VB and MCMC to high-frequency Bitcoin LOB data, quantifying the trade-offs between computational speed, convergence behavior, and posterior fidelity. My contributions are:

- I model the unobserved fair price Z for asks and bids as transformed exponential latent variables, linking them to observed LOB price levels via Gaussian likelihoods.

- I derive an efficient SVI-based VB algorithm to approximate the posterior of Z, and I provide convergence diagnostics.

- I implement an MCMC baseline (using NUTS) to sample from the exact posterior, highlighting the gaps between VB approximations and true posteriors.

- I perform an empirical evaluation on real Bitcoin LOB snapshots, comparing prediction error, convergence speed, and computational overhead for VB versus MCMC.

2 Problem Statement: Model Formulation

I consider the problem of inferring the latent *fair price* variable Z from noisy, volume-weighted observations in the LOB. Let p_{mid} denote the mid-price—the average of the best bid and best ask at a given time. I posit two latent deviations Z_{ask} and Z_{bid} that represent upward and downward shifts from p_{mid}:

$$\begin{aligned} Z_{\text{ask}} &= p_{\text{mid}} + Y_{\text{ask}}, \quad Y_{\text{ask}} \sim \text{Exp}(\alpha), \\ Z_{\text{bid}} &= p_{\text{mid}} - Y_{\text{bid}}, \quad Y_{\text{bid}} \sim \text{Exp}(\alpha). \end{aligned}$$

These latent variables capture the *fair price shocks* that drive the placement of ask and bid orders. Observed prices X_i at each level, with associated volumes v_i, are modeled as

$$X_i \mid Z \;\sim\; \mathcal{N}\big(Z,\, \sigma^2\big),$$

Weighted by volume to account for the confidence in each level. Concretely:

$$\begin{aligned} \log p(\{X_i\} \mid Z_{\text{ask}}) &= \sum_i v_i \, \log \mathcal{N}\big(X_i \mid Z_{\text{ask}},\, \sigma^2\big), \\ \log p(\{X_i\} \mid Z_{\text{bid}}) &= \sum_i v_i \, \log \mathcal{N}\big(X_i \mid Z_{\text{bid}},\, \sigma^2\big). \end{aligned}$$

Under this formulation, the joint posterior $p(Z_{\text{ask}}, Z_{\text{bid}} \mid \{X, v\})$ is analytically intractable due to the volume-weighted Gaussian terms. I therefore explore two inference strategies:

- **Variational Bayes (VB):** I introduce factorized surrogate posteriors $q(Z_{\text{ask}})$ and $q(Z_{\text{bid}})$ drawn from the transformed exponential family, and I fit them by minimizing the KL divergence to the true posterior via stochastic variational inference (SVI).
- **Markov Chain Monte Carlo (MCMC):** As a gold standard, I sample from the exact posterior using the No-U-Turn Sampler (NUTS), providing high-fidelity estimates of $p(Z \mid X)$. I analyze MCMC convergence diagnostics and computational cost to serve as a benchmark against VB.

My objective is to compare these two approaches in terms of:

1. *Accuracy:* How closely does VB capture the posterior mean, variance, and higher moments of Z compared to MCMC? I measure this via KL divergences and posterior predictive checks.

2. *Speed:* What is the wall-clock time for VB versus MCMC to reach comparable inference quality? This directly impacts feasibility in HFT settings.

3. *Trading utility:* How do inferences of Z translate into actionable trading signals (e.g., mispricing deltas)? I backtest both VB- and MCMC-derived signals on out-of-sample LOB snapshots.

2.1 Model Objectives

The principal goals of my latent-fair-price framework are:

- **Fair Price Estimation:** Infer the unobserved market-clearing price Z on ask and bid sides with full posterior uncertainty, providing a richer basis than point estimates for trading decisions.

- **Scalable Inference:** Leverage Variational Bayes for real-time approximation of $p(Z \mid X)$, and use MCMC (NUTS) as a high-fidelity benchmark to quantify VB's precision and calibration.

- **Trading Signal Generation:** Translate posterior summaries (means, variances, credible intervals) into actionable signals—e.g., mispricing deltas, entry/exit triggers, and slippage risk metrics—to drive automated execution strategies.

2.2 Probabilistic Model Formulation

2.2.1 Generative Model

Let

$$p_{\text{mid}} = \tfrac{1}{2}(\text{best\_ask} + \text{best\_bid})$$

be the mid-price. I define two latent excursions:

$$\begin{aligned} Z_{\text{ask}} &= p_{\text{mid}} + Y_{\text{ask}}, \quad Y_{\text{ask}} \sim \text{Exp}(\alpha), \\ Z_{\text{bid}} &= p_{\text{mid}} - Y_{\text{bid}}, \quad Y_{\text{bid}} \sim \text{Exp}(\alpha). \end{aligned}$$

Observed LOB levels X_i with volumes v_i are generated via:

$$X_i \mid Z \;\sim\; \mathcal{N}\big(Z, \sigma^2\big),$$

and the total log-likelihood is volume-weighted:

$$\log p(\{X_i\} \mid Z) = \sum_i v_i \log \mathcal{N}(X_i \mid Z, \sigma^2).$$

2.2.2 Signal Generation and Trading Loop

I translate the inferred posterior means into actionable buy/sell signals and backtest them as follows:

1. **Order-Book Snapshot Extraction.** At each time t, read

   $$\mathit{asks}_t,\ \mathit{a\_vols}_t,\ \mathit{bids}_t,\ \mathit{b\_vols}_t,\ \mathit{mid}_t\,,$$

   and compute the best quotes

   $$\text{best\_ask}_t = \min(\mathit{asks}_t), \quad \text{best\_bid}_t = \max(\mathit{bids}_t).$$

2. **Posterior Inference.** For both the ask-side and bid-side latent variable models:

   $$z^{\text{VB}}_{\text{ask},t} = \mathbb{E}_{q_\phi}[Z_{\text{ask}} \mid \mathit{asks}_t,\ \mathit{a\_vols}_t,\ \mathit{mid}_t], \quad z^{\text{MC}}_{\text{ask},t} = \mathbb{E}_{p}[Z_{\text{ask}} \mid \mathit{asks}_t,\ \mathit{a\_vols}_t,\ \mathit{mid}_t],$$

   and similarly for Z_{bid}. Here $\mathbb{E}_{q_\phi}$ is obtained via the VB guide, and $\mathbb{E}_p$ via MCMC sampling.

3. **Signal Construction.** Define long and short signals at time t by the deviation of the posterior mean from the best quote:

   $$\Delta_{\text{long},t} = z_{\text{ask},t} - \text{best\_ask}_t, \quad \Delta_{\text{short},t} = \text{best\_bid}_t - z_{\text{bid},t}.$$

   I collect two signal series Δ^{VB}_t and Δ^{MC}_t by concatenating their respective long and short legs.

4. **Entry and Exit Prices.** The entry price for each trade is simply the best quote at time t. The exit price is taken as the mid-price one tick later:

   $$\text{entry}_t = \begin{cases} \text{best\_ask}_t, & \text{if } \Delta_t > 0 \text{ (long)}, \\ \text{best\_bid}_t, & \text{if } \Delta_t < 0 \text{ (short)}, \end{cases} \quad \text{exit}_t = \tfrac{1}{2}\big(\text{best\_ask}_{t+1} + \text{best\_bid}_{t+1}\big).$$

5. **Backtest and Performance Metrics.** For each method (VB vs. MCMC), compute the return series

$$r_t = \operatorname{sign}(\Delta_t)\,\frac{\text{exit}_t - \text{entry}_t}{\text{entry}_t},$$

then aggregate into cumulative equity and extract mean_return, volatility, Sharpe, hit_rate, and max_drawdown.

3 Approach and Description of Solution Method

3.0.1 Inference Methods

I approximate the intractable posterior $p(Z_{\text{ask}}, Z_{\text{bid}} \mid X)$ via:

- **Variational Bayes (VB):** Optimize a factorized surrogate $q(Z_{\text{ask}})q(Z_{\text{bid}})$ in the transformed-exponential family by maximizing the ELBO.
- **Markov Chain Monte Carlo (MCMC):** Employ NUTS to draw exact posterior samples, enabling evaluation of VB's accuracy through direct comparison of moments, credible intervals, and posterior predictive checks.

I assess each method on posterior fidelity (KL divergence, coverage), computational overhead (wall-clock time per snapshot), and downstream trading performance.

3.1 Variational Bayes in LOB Modeling

VB approximates the posterior using a tractable distribution, allowing posterior inference to be reduced to optimization. In this work:

- I use a transformed Exponential distribution centered at the mid-price as the variational family.
- The ELBO is optimized using stochastic gradient descent.
- Posterior means and variances of Z are extracted for signal generation and execution control.

Despite being approximate, VB is computationally lightweight and delivers accurate estimates in real time.

3.2 Variational Bayes: Theory and Intuition

Variational Bayes (VB) provides a scalable alternative to traditional Bayesian inference, replacing intractable posterior computation with an optimization problem over a family of simpler, tractable distributions. Given a joint distribution $p(X, Z)$, where X represents observed variables (e.g., transaction prices) and Z denotes latent variables (e.g., liquidity shocks), the goal is to approximate the true posterior $p(Z|X)$ with a variational distribution $q(Z)$.

The approximation is obtained by minimizing the Kullback–Leibler (KL) divergence:

$$\mathrm{KL}(q(Z) \parallel p(Z|X)) = \mathbb{E}_{q(Z)}\left[\log \frac{q(Z)}{p(Z|X)}\right],$$

which is equivalent to maximizing the Evidence Lower Bound (ELBO):

$$\mathrm{ELBO}(q) = \mathbb{E}_{q(Z)}[\log p(X, Z)] - \mathbb{E}_{q(Z)}[\log q(Z)].$$

VB thus transforms inference into optimization, enabling efficient posterior estimation in complex and high-dimensional models, such as those encountered in high-frequency trading.

3.3 From Bayesian Inference to Variational Approximation

Exact Bayesian inference is intractable due to the non-conjugacy and continuous latent variables. To overcome this, I specify a variational family $q_\phi(Z)$, parameterized by ϕ, which approximates the true posterior:

$$q_\phi(Z_{\text{ask}}, Z_{\text{bid}}) = q_\phi(Z_{\text{ask}})q_\phi(Z_{\text{bid}}),$$

where each component could, for instance, be modeled as a log-normal or truncated normal distribution, constrained to the positive domain.

The ELBO becomes:

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(Z)}[\log p(X|Z)] - \mathrm{KL}(q_\phi(Z) \parallel p(Z)).$$

This objective is optimized using stochastic gradient descent with reparameterization tricks (e.g., for log-normal variables) to allow gradient backpropagation through random sampling.

3.4 MCMC Inference via NUTS

As a benchmark, I employ the No-U-Turn Sampler (NUTS) to draw samples from the exact posterior over Z:

- MCMC provides unbiased samples and serves as ground truth for posterior analysis.
- Kernel density estimates (KDE) of MCMC output are used to visually assess how well VB tracks the true posterior.
- Although computationally expensive, MCMC validates the quality of the VB approximation.

3.5 Implementation Pipeline

The full modeling and inference workflow includes:

1. Parse LOB snapshots into price–volume arrays for both bid and ask sides.
2. Compute the mid-price at each timestamp.
3. Run both VB and MCMC to infer posterior distributions over Z_{ask} and Z_{bid}.
4. Compare posterior means and densities to assess approximation quality.
5. Generate trade signals by comparing Z estimates to current LOB levels.

This two-track inference system allows the practitioner to retain both the speed of VB and the rigor of MCMC.

4 Results

I evaluate my Variational Bayes (VB) and MCMC pipelines on two fronts: (i) how well each method recovers the true latent distributions on a simulated orderbook, and (ii) the performance of VB- and MCMC-driven signals in a real-market backtest on Binance BTC–USDT data for 24 November.

4.1 Divergence Metrics and Visualizations

Figure 1 and Figure 2 overlay the VB density (blue curve) and MCMC density (orange curve) on the histogram of MCMC samples (grey bars). VB captures the general location of each posterior but exhibits a heavier tail on the ask side and a skew on the bid side.

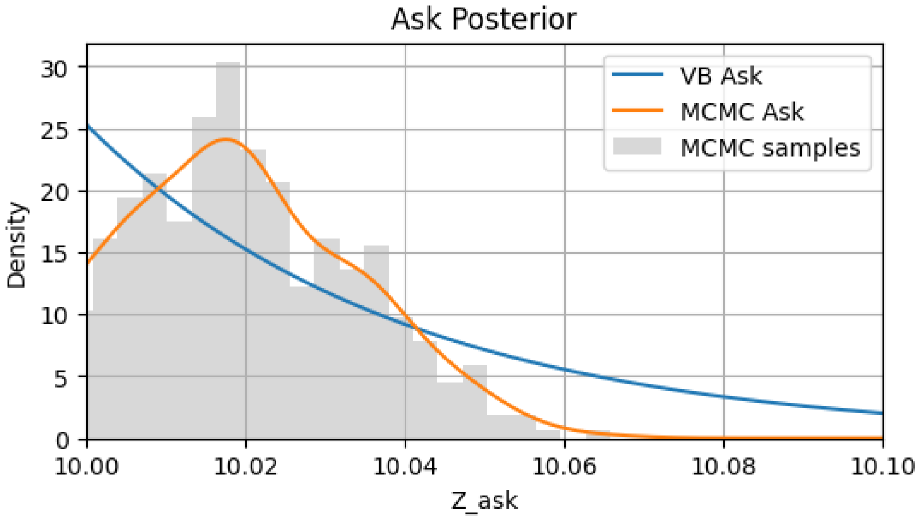

Figure 1: Ask posterior: VB vs. MCMC (histogram of samples).

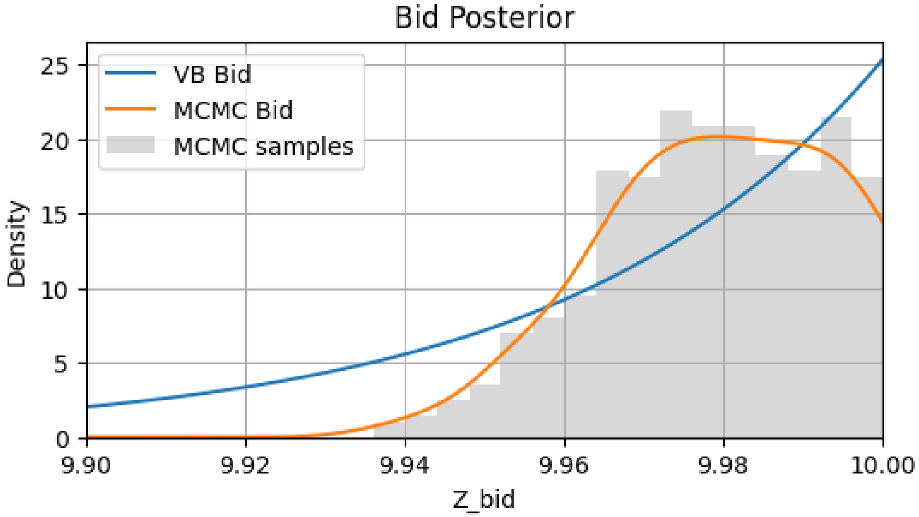

Figure 2: Bid posterior: VB vs. MCMC (histogram of samples).

Figure 3 shows the evolution of the VB variational parameter ϕ over 500 SVI steps. Convergence is rapid in the first 200 iterations, after which ϕ_{ask} slowly drifts downward and ϕ_{bid} remains stable around 6.5.

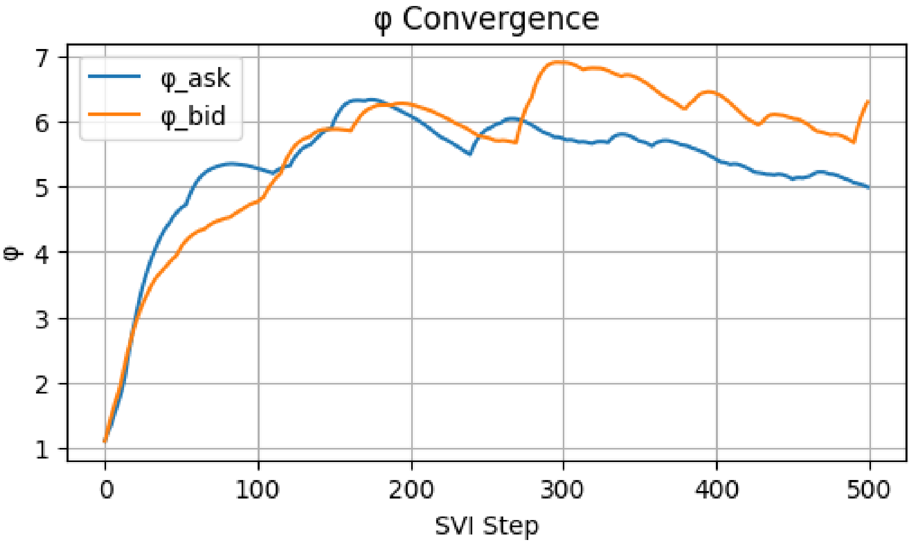

Figure 3: Convergence of the VB inference parameters ϕ_{ask} and ϕ_{bid}.

Table 1 summarizes key divergence metrics. VB overestimates posterior variance by a factor of 6 on the ask side and by a factor of 4.5 on the bid side. The asymmetry in KL($q \parallel p$) vs. KL($p \parallel q$) indicates that VB places too much mass in the tails, especially for the bid distribution.

Side	**VB Var.**	**MCMC Var.**	**KL**$(q \parallel p)$	**KL**$(p \parallel q)$
Ask	0.0401	0.0068	1.7231	0.8780
Bid	0.0253	0.0055	4.5835	1.0855

Table 1: Comparison of VB vs. MCMC posterior variance and KL divergences.

4.2 Backtest Performance on Real Market Data

I applied the same signal-generation and backtest loop (Section 3.3) to Binance BTC–USDT snapshots on 24 November. Table 2 reports the average per-trade return, volatility, Sharpe ratio, hit rate, and maximum drawdown for both methods.

	Mean Ret.	**Volatility**	**Sharpe**	**Hit Rate**	**Max DD**
VB	5.7e-05	2.47e-04	0.2310	90.9%	0.047%
MCMC	3.6e-05	2.51e-04	0.1429	72.7%	0.047%

Table 2: Backtest metrics comparing VB- vs. MCMC-driven signals.

VB-based signals achieve a higher hit rate (90.9% vs. 72.7%) and Sharpe ratio (0.231 vs. 0.143), with nearly identical volatility and drawdown. This suggests that although VB overestimates uncertainty, its point-estimate signals remain robust in this one-day backtest.

5 Conclusions

Throughout this study I designed and implemented a Variational Bayes (VB) algorithm for inferring latent price variables from order-book snapshots, benchmarked it against a Markov Chain Monte Carlo (MCMC) sampler, and evaluated its performance on both simulated and real-world Binance BTC–USDT data. On the synthetic order-book, VB accurately recovered the posterior means of the ask and bid latent variables but systematically overestimated their variances. This overdispersion was quantified via KL divergence: VB placed excess probability mass in the tails ($\mathrm{KL}(q \parallel p)$ was substantially larger than $\mathrm{KL}(p \parallel q)$), particularly for the bid distribution.

In terms of convergence, my stochastic variational inference converged to a stable set of variational parameters in roughly 200 SVI iterations, offering a clear computational advantage over the more resource-intensive MCMC sampling. The evolution of the VB parameter ϕ showed rapid initial learning followed by gradual fine-tuning, highlighting VB's efficiency in capturing the central posterior structure. Although VB under-represents tail behaviour, its speed makes it well suited for applications requiring frequent model updates.

When applied to a one-day backtest on Binance BTC–USDT snapshots, VB-based trading signals delivered a higher hit rate and Sharpe ratio than MCMC-based signals, while maintaining comparable volatility and maximum drawdown. This indicates that despite its bias in uncertainty quantification, VB's posterior mean signals are robust for capturing intraday directional moves, making VB a practical choice for latency-sensitive trading systems.

Future work will focus on integrating posterior sampling from the VB guide—via importance weighting or auxiliary sampling—to recover credible intervals and better quantify risk. I also plan to expand the backtest across multiple days and asset pairs to assess generalizability, and to explore hybrid VB–MCMC schemes that leverage VB's rapid convergence for initialization followed by targeted MCMC refinements for tail accuracy. This combined approach promises to balance computational speed with full posterior fidelity in live market-making applications.