

DYNAMICS OF LIQUIDITY SURFACES IN UNISWAP V3

JIMMY RISK, SHEN-NING TUNG, AND TAI-HO WANG

ABSTRACT. This paper presents a comprehensive study on the empirical dynamics of Uniswap v3 liquidity, which we model as a time-tick surface, $L_t(x)$. Using a combination of functional principal component analysis (FPCA) and dynamic factor methods, we analyze three distinct pools over multiple sample periods. Our findings offer three main contributions: a statistical characterization of automated market maker liquidity, an interpretable and portable basis for dimension reduction, and a robust analysis of liquidity dynamics using rolling window metrics. For the 5 bps pools, the leading empirical eigenfunctions explain the majority of cross-tick variation and remain stable, aligning closely with a low-order Legendre polynomial basis. This alignment provides a parsimonious and interpretable structure, similar to the dynamic Nelson-Siegel method for yield curves. The factor coefficients exhibit a time series structure well-captured by AR(1) models with clear GARCH-type heteroskedasticity and heavy-tailed innovations.

1. INTRODUCTION

Decentralized finance (DeFi) [HRS⁺21, GM23] is rapidly transforming financial markets, with automated market makers (AMMs) playing a crucial role in this evolution. As a revolutionary force, decentralized exchanges (DEXs) [CJ21] have emerged, attracting millions of users to DeFi projects and offering a way to bypass the constraints of traditional financial systems.

Introduced by Uniswap Labs, Uniswap marked a pivotal shift within the DEX landscape by replacing traditional order-book mechanisms with its innovative AMM design. Version 1 enabled censorship-resistant spot trading, allowing swaps between Ether (ETH), Ethereum’s native asset, and any ERC-20 token (e.g., USDC, WBTC). Version 2 [AZR20] generalized to any two ERC-20 tokens and added an on-chain oracle (a time-weighted average price). Version 3 [AZS⁺21] introduced concentrated liquidity, allowing liquidity providers (LPs) to allocate capital only within a chosen price interval. Fees accrue from swaps executed when the current price lies within the LP’s defined range.

Liquidity is a fundamental aspect of efficient markets, measuring the ease with which an asset can be exchanged without appreciable price impact [BBDG18]. In centralized venues, liquidity is extensively studied through limit-order books (LOBs); queueing and stochastic partial differential equation (SPDE) models [CST10, HLR15, CM21] effectively capture depth dynamics and resilience. By contrast, empirical work on AMM liquidity has so far centered on static range-allocation or fee-tier optimization. For example, Heimbach, Schertenleib, and Wattenhofer document that realized LP pay-offs hinge on tick-width and position lifetime [HSW22]. Fan et al. [FMCM⁺21] formalize the dynamic liquidity provision

Date: September 8, 2025.

Key words and phrases. Automatic market making, Decentralized exchange, Decentralized finance.

problem and focus on a general class of strategies to maximize LP earnings, and follow-up work provides a convex stochastic optimization problem that quantifies the optimal interval width for a belief-driven LP, providing implications for contract design [FMCA⁺22]. Related is work on optimal adaptive allocation through deep reinforcement learning [ZCY23], equilibrium liquidity distribution via mean-field games [BCN24], and predictable-loss-aware optimal strategies derived from dynamic continuous time models [CDM24].

On the more statistical side, panel regressions connect liquidity spreads and depth to gas costs, volatility, and fee revenue [ZLW⁺24]; researchers use linear models to forecast lookahead volume [MC23]; and statistical tests assess price discovery (across pools) and market efficiency [ACDF25]. Complementary work measures impermanent loss and LP risk [LHRW21] and clusters trader archetypes via graph embeddings [MC24].

In summary, liquidity plays a crucial role in understanding larger aspects of DeFi. A better understanding of the liquidity surface $L_t(x)$ aids comprehensive risk assessment, capital efficiency, and the development of sophisticated trading strategies. As mentioned, Uniswap v3 specifically generates a two-dimensional *liquidity surface* $L_t(x)$ indexed by block-time t and tick distance x that represents the quantized price range within the Uniswap v3 architecture. In other words, each trading pair has its own temporal landscape, enabling empirical liquidity analysis, inference, and modeling. Surprisingly, we find that a thorough statistical description of the $L_t(x)$ as a dynamic liquidity surface in (t, x) is still missing. This work is intended to fill that gap.

In particular, we apply methods in dynamic factor models and functional principal component analysis (FPCA) to obtain a low-rank factor representation. This has been applied to find low-rank structures in other dynamic functional financial objects, such as yield curves [Lit91, CDR11, Opr22], implied volatility surfaces [AHPP20], and forward curves [She09, HSH12].

To motivate this approach, consider the well-known Nelson–Siegel (NS) factor model as introduced by Charles Nelson and Andrew Siegel (1987) [NS87]. This was used as a parsimonious parametric framework to fit the term structure of interest rates (yield curves). The NS functional form uses three factors that are combined to approximate the cross-sectional shape of yields across maturities in a flexible but low-dimensional way [DLY08]:

$$\hat{y}_t(\tau) = \beta_{1,t} \cdot 1 + \beta_{2,t} \cdot \frac{1 - e^{-\lambda\tau}}{\lambda\tau} + \beta_{3,t} \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau} \right), \quad (1)$$

where where $\hat{y}_t(\tau)$ is the model-implied yield at time t and maturity τ , and $\beta_{1,t}$, $\beta_{2,t}$, $\beta_{3,t}$ are time-varying coefficients (factors) while λ is a decay parameter that governs the shape. While the original model [NS87] considered fixed coefficients (typically re-estimated), this *dynamic factor model* was a later extension by Diebold and Li (2006) [DL06] to incorporate time-varying factors. This extends the cross-sectional specification to time-varying factors with their own dynamics. Despite its simplicity, this Dynamic Nelson–Siegel (DNS) model provides an excellent fit to the term structure at each point in time and often yields forecasts that outperform more theoretically complex models in out-of-sample interest rate forecasting. By projecting the entire yield curve onto three persistent factors and then modelling those factors with low-order autoregressions, Diebold and Li achieved a practical and accurate forecasting method for yields [DL06]. Christensen et al. (2011) later connect with existing

no-arbitrage models for yield curves to add a no-arbitrage link to DNS without sacrificing parsimony [CDR11].

One of the remarkable aspects of the Nelson–Siegel basis is that its three latent factors naturally align with the principal components (PCs) observed in empirical yield curve data [DL06]. Most of the variation in the yield curve can be explained by these three orthogonal factors, colloquially known as level, slope, and curvature, which persistently match the factors in the NS model. Specifically, a principal component analysis on a panel of interest rates across maturities reveals the first PC to be a *level shift* affecting all maturities, the second PC a *slope factor* that tilts the curve by raising short rates and lowering long rates or vice versa, and the third PC a *curvature factor* that bends the curve (hump-shaped, affecting medium maturities more than short or long ones). This aligns with findings by Litterman and Scheinkman (1991) [Lit91], who, through eigendecomposition of the covariance matrix of zero-return vectors, showed that the first factor (level) explains 89–90% of variance and is nearly constant across maturities; the second factor (slope/twist) explains about 8% and exhibits a monotone sign change; and the third factor (curvature/butterfly) explains 1–2% and affects medium maturities more. This low-rank structure is also found in other financial surfaces. For instance, Cont and da Fonseca (2002) [CF02] observed a similar “level-slope-curvature” for implied volatility surfaces as functions of moneyness and time to maturity.

1.1. Contributions. In this study, we focus on the empirical dynamics of the liquidity surface $(t, x) \mapsto L_t(x)$ to answer three core questions: What is the statistical nature of these fluctuations? Can they be quantified? Can they be modelled in a parsimonious way? Our investigation begins nonparametrically, finding a stable low-rank structure when PCA is performed over various time windows. We also find that the nonparametric eigenmode basis consistently aligns well with the Legendre polynomial basis, providing an initial attempt at a persistent structure akin to the NS decomposition (1). By leveraging both decompositions, we identify explicit and persistent stochastic structures in the temporal components.

Our specific contributions are as follows:

- We provide a preliminary view of the surface $(t, x) \mapsto L_t(x)$ for three key liquidity pairs (Ethereum ETH-USDC 5bps (basis points), Ethereum ETH-USDC 30bps, and Arbitrum ARB-USDC 5bps) across three distinct time periods, accompanied by a high-level statistical analysis.
- We demonstrate that the infinite-dimensional dynamics (in x) can be effectively analyzed through a lower-dimensional principal component projection, yielding temporal coefficients and eigenmodes over tick distance.
- We identify persistent autoregressive [order 1] (AR(1))–generalized autoregressive conditional heteroskedasticity (GARCH) time series structures, with heavy-tailed innovations for the temporal coefficients.
- We show that the Ethereum ETH-USDC 5bps and Arbitrum ARB-USDC 5bps pools maintain a stable eigenmode structure that closely aligns with the Legendre polynomial basis, a conclusion we quantify using subspace distance metrics.
- We use the Legendre polynomials as an interpretable tool for the liquidity surface and illustrate the effect of its orthogonal components through volatility shocks.
- We verify the robustness of our statistical findings by replicating the analysis with various preprocessing schemes.

The rest of this work is outlined as follows: Section 2 covers the background, with 2.1 detailing the Uniswap v3 framework and the meaning of a liquidity surface, and 2.2 covering the statistical methods used. Section 3 provides details on the datasets and preprocessing used. The follow-up Section 4 performs the statistical analysis, including PCA and Legendre basis decompositions, assessing stability, persistent stochastic structures, and the effects of shocks. Lastly, Section 5 finalizes the discussion and provides avenues for future work.

2. BACKGROUND

2.1. Uniswap v3 Liquidity Provision. Uniswap v3 introduced a revolutionary approach to decentralized liquidity provision, allowing liquidity providers (LPs) to concentrate their capital within specific price ranges. This mechanism fundamentally redefines the relationship between an LP’s token holdings and their contribution to the market, moving beyond the uniform liquidity distribution of earlier Automated Market Maker (AMM) versions.

2.1.1. Concentrated Liquidity in Detail. Concentrated liquidity is the cornerstone of Uniswap v3’s capital efficiency. Unlike previous iterations, where liquidity was spread evenly across the entire price curve, Uniswap v3 empowers LPs to strategically allocate capital within a user-defined price range $[p_l, p_r]$. This innovative feature hinges on several interconnected concepts:

- (1) **Liquidity Level L :** This parameter quantifies the amount of liquidity an LP provides within their chosen range. Conceptually, it represents the virtual reserves contributed by the LP, analogous to the constant-product formula, $L = \sqrt{xy}$, used in Uniswap v2. An LP’s liquidity level directly determines its share of trading fees and its exposure to price movements as long as the market price remains within its defined range.
- (2) **Liquidity Provision Range $[p_l, p_r]$:** This is the specific price interval an LP selects for capital deployment. When the current market price p , as reported by the AMM, resides within this designated range, the LP’s assets become active for swaps, following a predetermined bonding curve [AZS⁺21]. The precise selection of this range is paramount, as it dictates the periods during which the LP’s capital actively participates in facilitating trades.
- (3) **Aggregated Liquidity:** Uniswap v3 achieves its distinctive *aggregated* liquidity profile by seamlessly combining contributions from numerous individual LPs, each providing liquidity within their own defined price ranges (see Figure 1). Denote the i -th LP’s position by its liquidity level L_i and its chosen price range $[p_{\ell,i}, p_{r,i}]$. An individual LP’s liquidity L_i is only active when the current market price p falls within their specified range. The total liquidity, $L_{\text{total}}(p)$, available in the Uniswap v3 pool at any given price p is then the sum of the liquidity levels of all individual LP positions whose active ranges encompass that price:

$$L_{\text{total}}(p) = \sum_{i=1}^N L_i 1_{[p_{\ell,i}, p_{r,i}]}(p).$$

This summation illustrates how the distinct liquidity contributions from individual LPs are effectively combined at each price point. This process creates a potentially

non-uniform liquidity distribution across the entire price spectrum, directly influenced by where LPs choose to concentrate their capital [TW24]. Consequently, an LP's token holdings and fee earnings are dynamically determined by its individual liquidity level L_i and how the market price p interacts with its active range $[p_{\ell,i}, p_{r,i})$.

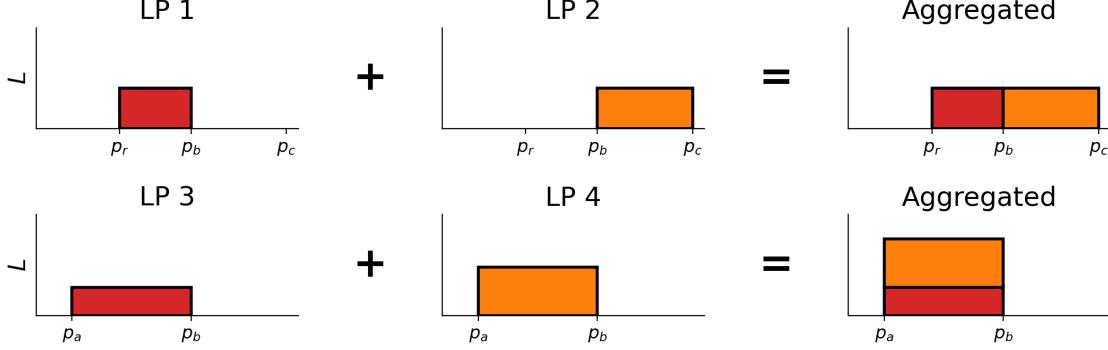


FIGURE 1. Aggregated Liquidity

2.1.2. Comparison with Traditional Limit Order Books. Uniswap v3 represents a significant evolution in AMM design, effectively bridging the gap between constant function market makers (CFMMs) [AC20] and traditional electronic limit order books (LOBs). While CFMMs like Uniswap v2 offer simplicity, they lack the expressiveness of LOBs, which maintain a list of outstanding buy and sell orders at specified prices [GPW⁺13]. Uniswap v3's concentrated liquidity feature allows liquidity provider (LP) positions to function akin to limit orders in LOBs, thereby approximating more complex demand curves without the full state complexity of a traditional order book [MMR23].

Despite sharing the goal of enabling trades at preferred price levels, Uniswap v3's narrow-range LP positions and traditional LOB limit orders exhibit fundamental differences in their underlying mechanisms and operational characteristics. These distinctions are detailed below:

- (1) **Price Specification:** A primary distinction lies in how the price is specified. In Uniswap v3, LPs define a price *range*, providing continuous liquidity across that interval. Conversely, LOBs rely on specific, *discrete price points* for each individual order.
- (2) **Liquidity Provision Mechanism:** Liquidity provision differs fundamentally. Within a defined range, Uniswap v3 liquidity is continuous and formula-driven, always available according to the bonding curve until fully utilized. In contrast, LOB liquidity is offered in discrete, fixed quantities at specific price levels.
- (3) **Operational Mechanisms and Trade Execution:** Uniswap v3 operates as an AMM, utilizing a constant-product formula (adjusted for the specified price range) to facilitate swaps. Trades execute instantly against the available liquidity within an LP's range until that liquidity is depleted. In contrast, LOBs function by matching submitted buy and sell orders, executed through a greedy matching process at the specified price or better.

- (4) **Behavior with Market Price Movements:** The behavior of positions under market price fluctuations also diverges. If the market price moves outside of a Uniswap v3 LP's defined range, the position becomes inactive, holding its current asset composition until the price re-enters the range. This often necessitates active management and potential position withdrawal by the LP to realize intended gains or avoid impermanent loss. In LOBs, limit orders remain in the book until they are either filled or explicitly cancelled by the user, irrespective of broader market price movements.

2.2. Statistical Methods. Suppose for $x \in \mathcal{X} = [-1, 1]$ and $t \geq 0$ we observe $\{y_t(x)\}$. To connect with the previous discussion, $y_t(x) = \log L_t(x)$, and tick x has been standardized to $[-1, 1]$ (see Section 3.2.1 for details). The idea is to decompose

$$y_t(x) = m(x) + \sum_{k=1}^K \beta_{t,k} u_k(x) + r_t^{(K)}(x) \quad (2)$$

into K *factors* with *basis* (or loadings) $\{u_k(x)\}_{k=1}^K$ and *coefficients* (or factor scores) $\{\beta_{t,k}\}_{k=1}^K$, and $r_t^{(K)}(x)$ is the *residual truncation error*. Through this decoupling, one can separately analyze the temporal dynamics through the time series $(\beta_{t,k})_{t \geq 0}$ and spatial structure through the $(u_k(x))_{x \in \mathcal{X}}$. Loosely, the goals are

- K small enough and an interpretable structure on the $\{u_k\}_{k=1}^K$ so that the model remains parsimonious and interpretable;
- some persistence so that certain properties (e.g., stable basis, distributions of coefficient processes) are stable across time windows;
- K large enough for an appropriate basis $\{u_k\}_{k=1}^K$ so that $r_t^{(K)}(x)$ is (approximately) idiosyncratic.

It is important to note that the first and last goals are often difficult to achieve simultaneously.

2.2.1. Dynamic Factor Extraction. In practice, we observe a data matrix Y with entries $\{y_t(x_m)\}$ for $t = 1, \dots, T$ and $m = 1, \dots, M$. We assume the x_m are equally spaced. Each row of Y corresponds to the observed time- t curve, $\mathbf{y}_t = [y_t(x_1), \dots, y_t(x_M)]$. Given the lack of prior information about the statistical and functional dynamics of $y_t(x)$, a nonparametric approach such as principal component analysis (PCA) is appropriate. For this reason, we assume the data are centered unless otherwise noted; otherwise, replace \mathbf{y}_t with $\mathbf{y}_t - \bar{\mathbf{y}}$ where $\bar{\mathbf{y}} = \frac{1}{T} \sum_{t=1}^T \mathbf{y}_t$.

The following approach applies dynamic factor models [SW11], which are both conceptually and mathematically similar to functional principal component analysis (FPCA) [RS05]. The former is often employed in econometric settings where there is no necessary structure on the M coordinates. The latter considers all objects in $L^2(\mathcal{X})$ rather than \mathbb{R}^M , thus imposing structure in the x coordinate. Our equally-spaced and dense setup recovers the same subspace as FPCA in the limit; see Section A for details. Note that despite having an enticing name, *dynamic FPCA* as popularized by Hormann et al. [HKH15] is not appropriate as it replaces (2) with a non-factorable double sum (one infinite), improving theoretical properties but hindering the purpose of parsimonious interpretability.

Denote the sample $M \times M$ covariance matrix as

$$\hat{\Sigma} := \frac{1}{T} Y^\top Y = U \Lambda U^\top, \quad (3)$$

where $U \Lambda U^\top$ is its eigendecomposition (principal component) form. Here, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_M)$ are the eigenvalues of $\hat{\Sigma}$ and U has orthonormal columns consisting of $\hat{\Sigma}$'s orthonormal eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_M$ with $\mathbf{u}_k = [u_k(x_1), \dots, u_k(x_M)]^\top$. For simplicity, assume $T > M$ and full rank, so that the eigenvalues can be ordered $\lambda_1 > \lambda_2 > \dots > \lambda_M > 0$.

For a fixed $K \leq M$, PCA yields the optimal rank- K least-squares reconstruction:

$$\{\beta_{t,k}\}, U_K = \underset{\{\tilde{\beta}_{t,k}\}, \tilde{U}_K}{\text{argmin}} \sum_{t=1}^T \sum_{m=1}^M \left(y_t(x_m) - \sum_{k=1}^K \tilde{\beta}_{t,k} \tilde{u}_k(x_m) \right)^2, \quad (4)$$

where $\beta_{t,k} = \langle \mathbf{y}_t, \mathbf{u}_k \rangle$ under the Euclidean inner product on \mathbb{R}^M . This solution is also unique (up to sign) if we assume ordered eigenvalues and orthonormality on the U_K . If \mathbf{y}_t is weakly stationary with $\text{var}(\mathbf{y}_t) = \Sigma = U \Lambda U^\top$, then the same holds true at the population level:

$$\{\beta_{t,k}\}, U_K = \underset{\{\tilde{\beta}_{t,k}\}, \tilde{U}_K}{\text{argmin}} \mathbb{E} \left\| \mathbf{y}_t - \sum_{k=1}^K \tilde{\beta}_{t,k} \tilde{\mathbf{u}}_k \right\|^2. \quad (5)$$

In summary, the optimal rank- K representation in (2) is the one produced by the covariance (according to squared error, and under expectation in the population setting).

Remark: If $\{y_t(x)\}$ is weakly stationary, $\hat{\Sigma}$ is a consistent estimator of the population covariance. If it is not, the decomposition and optimality of (3) and (4) still hold, but the interpretations change (see Section 3.2.1).

We shift the analysis to the truncated

$$y_t^{(K)}(x_m) = \sum_{k=1}^K \beta_{t,k} u_k(x_m), \quad m = 1, \dots, M \quad (6)$$

so that t and x have been decoupled. The interpretation is that each k th *factor* consists of the pair $(\beta_{t,k}, u_k(x))$, and is one piece that contributes orthogonally to $y_t^{(K)}(x)$, in order of decreasing variance explained. Often one chooses K via the cumulative proportion of variance explained, where the (cumulative) proportion of variance explained by factor k is given by

$$\text{PVE}_j = \frac{\lambda_j}{\sum_{j=1}^M \lambda_j}, \quad \text{CPVE}_K = \sum_{j=1}^K \text{PVE}_j, \quad (7)$$

so that $\text{CPVE}_M = 1$. If one finds that K is suitable to achieve a desired threshold (say, 95%), these K components are individually analyzed. The $u_k(x)$ offer insight into a static structure of $y_t(x)$ in the x coordinate. Under weak stationarity, these are time-invariant (asymptotically in the sample case). The scores $\{\beta_{t,k}\}$ capture temporal dynamics (e.g. AR(1) for k or vector AR(1) (VAR(1)) across k). Statistical properties of the $\beta_{t,k}$ transfer to the associated structural effect $u_k(x)$ when reconstructing $y_t(x)$. Thus, the *stochastic properties of the k th factor*, such as mean reversion, heavy tails, or heteroskedasticity, can be analyzed through its corresponding $\beta_{t,k}$.

While PCA provides a data-driven basis, we also explore the alignment of this basis with a fixed, interpretable basis, such as the Legendre polynomials. Details on these comparisons and the methodology for evaluating off-grid points are provided in Appendix A.

2.2.2. Forecasting. One obtains forecasts naturally by forecasting the $\beta_{t,k}$ through traditional time series methods (e.g. AR(1) or VAR(1) in discrete time). Let $\mathcal{F}_t = \sigma(\{y_s(x_i)\}_{i=1}^M, s \leq t)$ be the *filtration* containing all relevant information up to time t and assume the loadings $u_k(x) \in \mathcal{F}_t$, either estimated and fixed or by using a fixed basis. Define

$$\hat{\beta}_{t+h|t,k} := \mathbb{E}[\beta_{t+h,k} | \mathcal{F}_t], \quad \hat{y}_{t+h|t}^{(K)}(x) := \mathbb{E}[y_{t+h}^{(K)}(x) | \mathcal{F}_t].$$

Then by linearity

$$\hat{y}_{t+h|t}^{(K)}(x) = \sum_{k=1}^K \hat{\beta}_{t+h|t,k} u_k(x).$$

Further, $\hat{y}_{t+h|t}(x) = \hat{y}_{t+h|t}^{(K)}(x)$ if we assume the remainder satisfies $\mathbb{E}[r_{t+h}^{(K)}(x)|\mathcal{F}_t] = 0$. Consequently, prediction intervals and other distributional quantities can be obtained. Note that basis drift affects optimal forecasts, but it is usually mild for short horizons. In practice, the basis is periodically updated.

3. DATA AND METHODOLOGY

This section details the specific Uniswap v3 liquidity data used in this study, the necessary preprocessing steps, and how the data's characteristics inform our modelling approach. Our primary focus is on understanding liquidity behavior relative to the current market price through in-sample examination.

3.1. Uniswap v3 Liquidity Data Characteristics. Our analysis begins with raw liquidity data sourced from an internal API provided by Teahouse Finance, a decentralized finance (DeFi) asset-management platform specializing in Uniswap. The goal is to understand the aggregated liquidity dynamics over time and the discretized price space. Seminal work in Uniswap v3 provides some preliminary stylized facts pertinent to our modelling approach [FMCM⁺21, FMCA⁺22]:

- A large fraction of the liquidity mass lies near the current price, meaning we typically observe $L_t(x)$ to be relatively large near $x = 0$. Here, x represents the normalized logarithmic pool price, with $x = 0$ corresponding to the current pool price. Note that there may still be peaks further out.
- The shape of the liquidity surface can vary significantly depending on LP risk preferences and chosen price interval configurations.
- Market volatility significantly influences LP behavior. In high-volatility regimes, LPs tend to widen their liquidity ranges, which shifts liquidity mass outwards and consequently lowers the central peak. Conversely, periods of low volatility are often characterized by a tighter, steeper concentration of liquidity around the mid-price.

One goal is to verify these to the best of our ability and explain them through the dynamic factor lens. Throughout, we add additional stylized facts that arise through the statistical analysis.

3.1.1. Datasets Utilized. Our analysis incorporates data from a variety of Uniswap v3 pools to assess commonalities, generalizability, and improve robustness of takeaways:

- **Ethereum Mainnet 5bps ETH-USDC pool:** This pool serves as our primary focus due to its consistently high trading volume and deep liquidity, providing a rich dataset for in-depth analysis.
- **Ethereum Mainnet 30bps ETH-USDC pool:** Included to investigate how different fee tiers (and thus different ‘tickSpacing’ values) impact liquidity distribution and dynamics, allowing for a comparison against the 5bps pool.
- **Arbitrum 5bps ARB-USDC pool:** Included as a crucial robustness check to assess if observed features generalize across distinct blockchain environments, specifically Arbitrum (an Ethereum Layer-2 scaling solution), which has different block times and transaction costs.

The times and block numbers are displayed in Table 1. Block spacing was chosen as 115200 for ARB5 and 2400 for ETH to obtain an (approximate) 8-hour time spacing.

Dataset	Pool	bps	Time Range			Spacing
			Start	End		
ARB5	ARB-USDC	5	Date-time: Block Number:	Aug 23, 2023 124163200	Nov 26, 2024 278646400	8.39 h 115200
ETH5, ETH30	ETH-USDC	5, 30	Date-time: Block Number:	Aug 1, 2021 12940529	Nov 26, 2024 21274529	8.26 h 2400

TABLE 1. Full dataset ranges. Dates are rounded to calendar days (UTC). The Spacing column reports time spacing (hours) on the date row and block step on the block row. ETH5 and ETH30 share the same blocks and times. Spacing in hours is computed as the average hourly time between the clock times at the exact block-number intervals.

3.2. Applying Dynamic Factor / Functional PCA Models. We give details on our implementation of Uniswap v3 liquidity into the framework in Section 2.2. Time t is *block number*, a discrete index advanced by consensus when a new block is appended to the chain. Block numbers provide a consistent, immutable, and sequential temporal proxy for ordering events. In general, t is stochastically related to its corresponding clock time $c(t)$. For example, Ethereum is partitioned into slots of 12 seconds, and at most one block can be proposed per slot. One typically observes $c(t) - c(t-1)$ to be approximately 12 seconds, but it can sometimes be greater due to skipped or empty blocks and other factors [KM23]. Arbitrum is slightly different and has an approximate gap of 0.25 seconds per block [KGC⁺18]. Thus, the liquidity surface can be indexed by either the discrete block time t or clock-time $c(t)$. When there is no ambiguity, we may interchange t or $c(t)$ for clarity.

Next, Uniswap v3’s price space is discretized into *ticks*. The price at an integer tick index i is defined by $P(i) = 1.0001^i$, meaning each tick represents a 0.01% (1 basis point) price movement. Liquidity ranges for LPs are set between specific *tick-spacing multiples* s . The parameter s varies by pool fee tier (e.g., a 0.05% fee tier uses $s = 10$) and directly determines the minimum width and granularity of liquidity provision ranges. The total liquidity available at any given price p is the sum of individual LP contributions whose active ranges encompass

that price, as previously defined in Section 2.1.1. This aggregation results in the non-uniform liquidity distribution we aim to model.

To analyze liquidity dynamics relative to the current market price, we employ a relative tick coordinate system, analogous to how liquidity in LOBs is often described relative to the best bid/ask [CM21]. Essentially, this normalizes the current pool price tick to 0, which aids in capturing the functional form of liquidity, where behavior is assumed to be more consistently described when referenced to the current AMM price rather than absolute price levels. Given the vast range of possible Uniswap v3 ticks, this transformation simplifies analysis by focusing on relative behavior. For background on price variations in financial markets, see [Con01].

To simplify our modelling approach, we analyze the liquidity at a fixed number of relative ticks around the current market price (See Appendix B for details). This method allows us to focus on the shape of the liquidity surface close to the price, effectively transforming a variable-length dataset into a fixed-length functional form suitable for analysis.

3.2.1. Further Preprocessing. We make a few miscellaneous comments about preprocessing in the subsequent analysis.

- We work with the logarithmically transformed $y_t(x) = \log L_t(x)$ as it matches with how x is related to the log price (since $\log P(i) = i \log 1.0001$), and also helps stabilize the associated probability distribution. Furthermore, this is common practice in the analysis of financial curves to ensure positivity.
- We work with the undifferenced log-liquidity data. This is common in situations where interpretable factors are paramount. For example, the dynamic Nelson-Siegel framework for yield curves is applied to undifferenced data to directly capture the level, slope, and curvature of the term structure [DL06], despite unit root behavior in many samples. Here, the focus is on providing an interpretation of the inherent shape and evolution of the liquidity surface rather than time differences, obtaining an interpretable and parsimonious decomposition. Evidence and implications of nonstationarity are discussed throughout the analysis.

4. EMPIRICAL ANALYSIS

This section presents the empirical analysis of Uniswap v3 liquidity surfaces. The analysis is broken into two main types. The first uses the full data in Table 1 and works over overlapping rolling windows to assess how certain phenomena persist in time. The second focuses on three specific windows with a more in-depth statistical analysis and interpretation in each, comparing across windows and datasets. The windows are described in Table 2. These windows serve to localize the principal component decomposition to illustrate how within-window results provide valuable insights and are aligned across datasets (up to block-number variability).

Hyperparameter choices. The main analysis utilizes $M = 201$ ticks, so $x = 0$ is the current price and $x \in \{-1.0, \dots, -0.01\}$ and $\{0.01, \dots, 1.0\}$ are the 100 (log) prices on either side. Additionally, the data are provided using `blockNumber` as the time unit. We use a block-number spacing so that $y_t(x)$ and $y_{t-1}(x)$ are approximately 8 hours apart; this corresponds to 115200 blocks for ARB5 and 2400 for ETH. This analogously matches a typical 8-hour

Window		Arbitrum ARB-USDC		Ethereum Mainnet ETH-USDC	
		Start	End	Start	End
1	Date	Sep 1, 2023	Jan 22, 2024	Sep 1, 2023	Jan 13, 2024
	Block	126928000	173008000	18038129	18998129
2	Date	Feb 1, 2024	Jun 16, 2024	Feb 1, 2024	Jun 14, 2024
	Block	176233600	222313600	19130129	20090129
3	Date	Jul 1, 2024	Nov 12, 2024	Jul 1, 2024	Nov 12, 2024
	Block	227497600	273577600	20210129	21170129

TABLE 2. Training windows for each dataset. Each window contains $T = 400$ observations spaced at approximately 8-hour intervals. Dates are rounded to calendar days (UTC). Ethereum 5bps and 30bps share the same windows and blocks.

daily trading window in centralized exchanges. We initially tried a 2-hour spacing, but found that movement in the liquidity surface was too sparse to model reliably. Note that liquidity data are updated continuously, so there are no breaks during nighttime or during holidays. Additionally, whenever PCA is performed, it uses $T = 400$ rows of data (approximately 4.5 months), which provides enough data for analysis but not too long a window to over-generalize.

Appendix F performs analysis of alternative choices of M and also $T = 800$. Comparative details are discussed in the main text when relevant.

4.1. Initial Inspection of Log-Liquidity Surface. As an initial inspection, Figure 2 displays a 3D surface plot of $(t, x) \mapsto y_t(x)$ (top panel) for the second window of each dataset, and a fixed-time cross-sectional plot of $y_{t_0}(x)$ vs x , where t_0 is the first time in each window (bottom panel). The first and third windows are left out of the top panel for space reasons.

The 3D surfaces generally show liquidity to be unimodal, though the mode may shift over time, and additional temporary modes can appear due to stochastic fluctuations. This is clearly present in the snapshots for ARB5 and ETH30, and both appear roughly concentrated around the current market price ($x = 0$). ARB5 has some asymmetric tilting across time points, and ETH30 has shifts in variability, with the latest date displayed (Jul 1, 2024) having noticeably lower values toward the edges, and some spikes toward the center. The ETH5 pool exhibits a flatter profile (top panel) and has several spikes in its cross-sections, making the general pattern less discernible. These irregularities are also visible in its surface plots, with some also appearing in ARB5 at various points.

To better detect how liquidity profiles shift within a 4.5-month period, we compute the sample mean and standard deviation functions of the log-liquidity:

$$\bar{y}(x) = \frac{1}{T} \sum_{t=1}^T y_t(x), \quad s(x) = \sqrt{s^2(x)}, \quad s^2(x) = \frac{1}{T} \sum_{t=1}^T (y_t(x) - \bar{y}(x))^2 \quad (8)$$

over each window and also the full dataset. These are plotted in Figure 3.

Consistent with Figure 2, all mean functions in Figure 3 display a tent-like shape centered at $x = 0$, generally consistent with the mean over the full data. ARB5 and ETH5 have

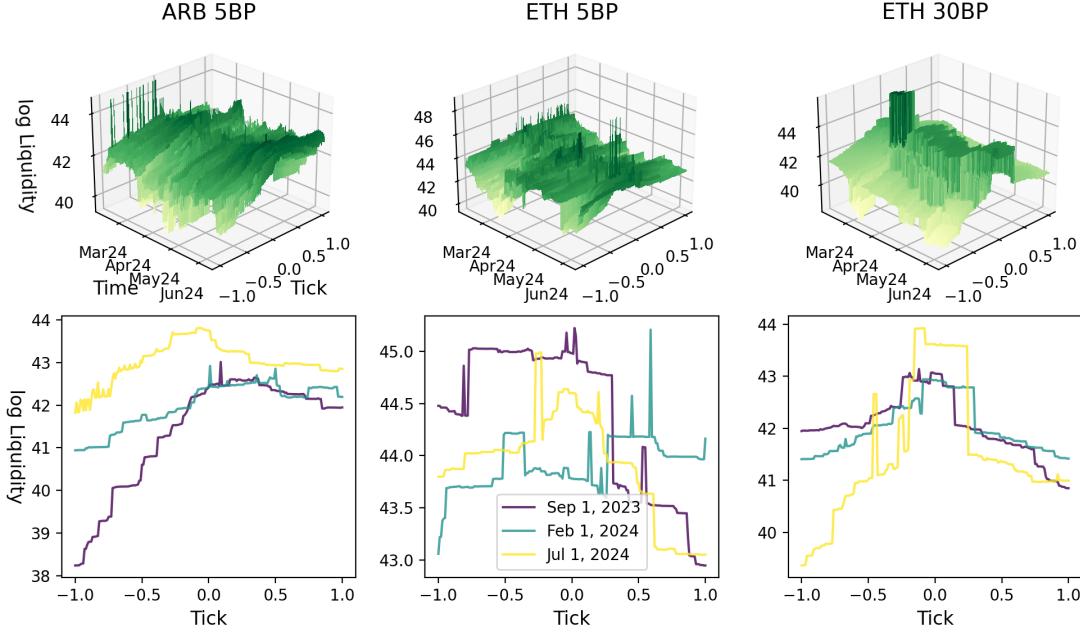


FIGURE 2. Top panel: raw log-liquidity surfaces $y_t(x)$ over time t and relative tick x according to Window 2. Bottom: cross-sectional plots of $y_{t_0}(x)$ against x with t_0 being the first time in its respective window.

notably stable shapes in the mean (although there are shifts), with some flattening occurring in Window 3 for both (more exaggerated for ETH5). ETH30 has a similar shape for the first two windows (and full average), but has a less discernible ‘tent’ shape in the third window, with a more flattened mode whose location is shifted toward $x = -0.25$.

The standard deviation functions of the three datasets vary significantly. While ARB5 and ETH5 share some similarities, such as a valley shape (which can be tilted or smoothed) with higher variance at the edges, they also show distinct characteristics. The ETH5 dataset, in particular, has a spike in variance at $x = 0$, which then rapidly drops before increasing again towards the edges. The asymmetry in variance for both ARB5 and ETH5, with higher variability around negative ticks, may be due to LPs anticipating price movements. In contrast, the ETH30 dataset exhibits a noticeably different and less stable structure. This is evidenced by the lack of a clear, persistent ‘tent’ shape and significant shifts in its standard deviation function over different time periods, reflecting more complex liquidity dynamics in this higher-fee pool.

4.2. Persistence of Variation Explained. With the noticeably different shapes across windows, we perform rolling window analysis to see how eigenvalues and proportion of variance explained vary across time. To clarify the timing and meaning of *rolling window*, we denote $t_j = 0, 10, 20, \dots$ for $j = 1, 2, \dots$ as the start dates of a rolling window, where $t_j = 0$ is the same as $t = 0$ for the full dataset in Table 1. Then PCA is performed over the rolling liquidity surface $(y_{t_j}(x), y_{t_j+1}(x), \dots, y_{t_j+T-1}(x))$ for each $j = 1, 2, \dots$. Thus, each contains its own decomposition and summary statistics. As before, $T = 400$ and $M = 201$ so that

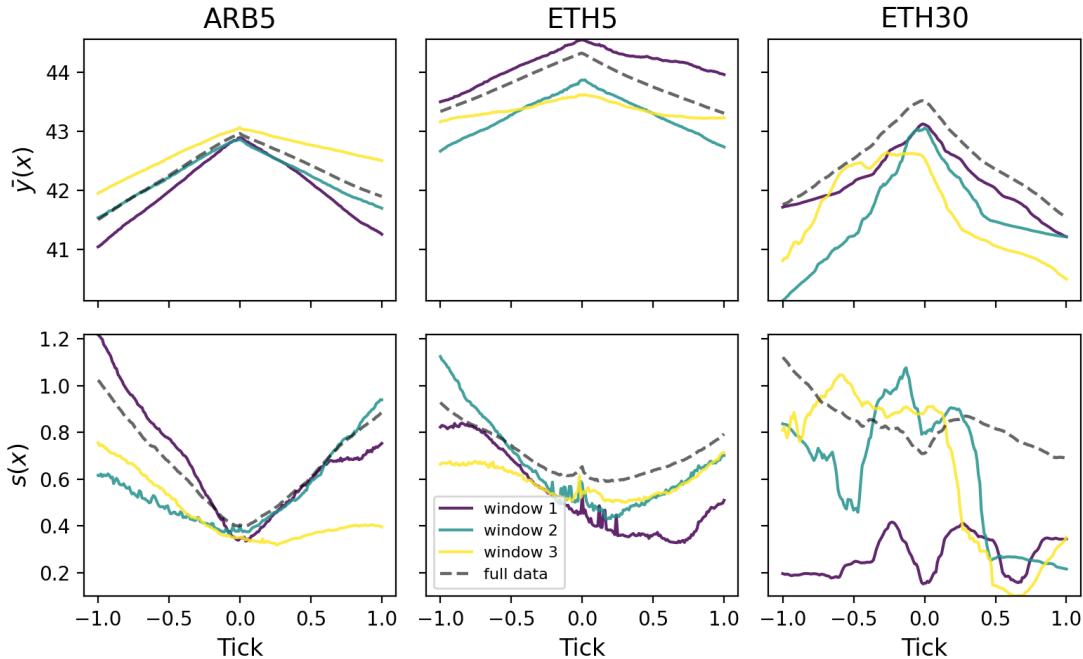


FIGURE 3. Sample mean and standard deviation functions taken over its corresponding window, plotted against relative tick x , across all datasets and windows in Table 2. The dashed black lines are averages taken over the full dataset in Table 1.

each surface shares $400 - 10 = 390$ times with the previous in sequence. Useful summary statistics over the windows are shown in Figure 4.

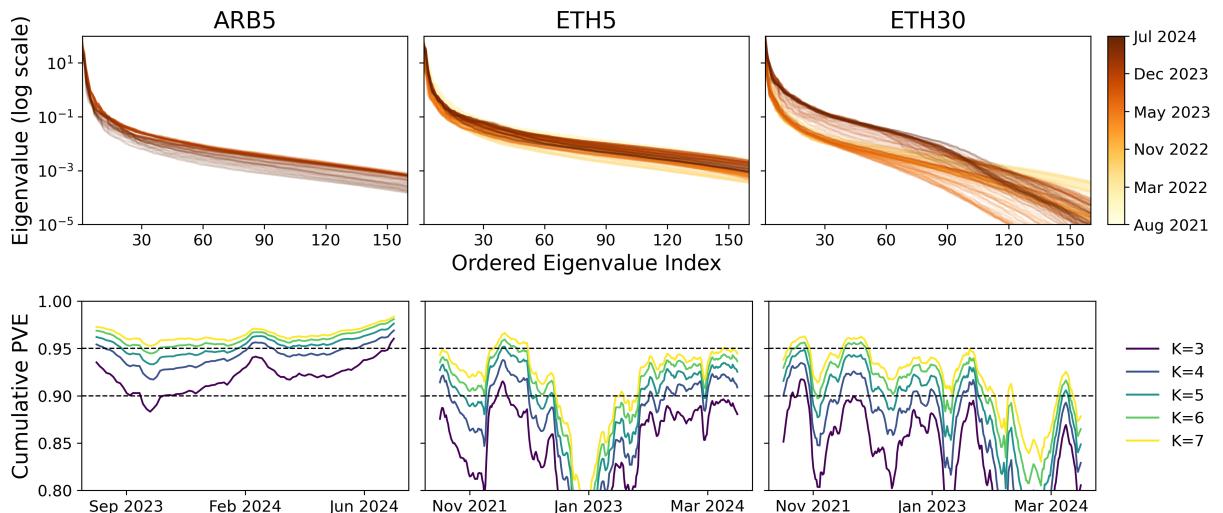


FIGURE 4. Rolling-window principal component analysis (PCA) of liquidity surfaces. Top: ordered eigenvalues (log-scale) for each window of length $T = 400$, color-coded by the window start date t_j (lighter = earlier, darker = later). Bottom: cumulative proportion of variance explained (CPVE; Eq. (7)) versus t_j for fixed ranks $K \in \{3, 4, 5, 6, 7\}$. Columns correspond to ARB5, ETH5, and ETH30.

The top panel shows the ordered eigenvalues on the log-scale, colored by t_j . The lighter colors (toward yellow) are at earlier dates, and the darker (brown) are later. All three pairs experience a rapid decay up to the 25th or 30th eigenvalue. ARB5 and ETH5 consistently stabilize as linear afterward, with stable slopes and slightly shifted intercepts over time periods. Moving to ETH30, it has a similar shape to the other two prior to 2023 (yellow and light orange), but exhibits a persistent regime shift in late 2023, with a slower decay both in the higher ranks and later ranks (up to approximately the 90th rank). The stability across dates suggests that a lower rank structure is suitable for ARB5 and ETH5, while ETH30 may be too unstable.

The bottom panel of the figure looks specifically at the proportion of variance explained over the rolling windows for ranks $K = 3, 4, 5, 6, 7$. The y -axis is the cumulative PVE as in Equation (7) and the x -axis is the rolling window start date. A rank-5 structure appears adequate for ARB5 to provide 95% of variance explained, nearly uniform over the times considered; a rank 3 structure satisfies 90% cumulative PVE. ETH5 has a similar conclusion for $K = 5$ aside from a period of high volatility in Summer 2022 to Summer 2023. ETH30 is generally more volatile, requiring five components to explain 90% of variance up to Summer 2023; seven is generally not adequate up to 95%. Additionally, the regime shift observed in the top panel (late 2023) is present for ETH30: its cumulative PVE drops but does not stabilize as of the end date of data collection.

In summary, ARB5 and ETH5 are generally stable in eigenvalue decay. ARB5 and ETH5 share many similarities in their cumulative PVE, although ETH5 experiences a drift to require more PCs to attain a certain PVE around Jan 2023. Note that it is possible that ARB5 could also experience this phenomenon, but it is unobservable due to a lack of data. ETH30 has temporal irregularities across both panels.

Appendix F provides PVE plots for the choices of $M = 101, 51$, and 11 ticks in comparison to $M = 201$; see Figure 20. There is little difference in moving from $M = 101$ to $M = 201$, which justifies the choice of $M = 201$, providing a balance between a sufficient and necessary amount of data.

4.3. PCA: Basis Functions And Their Interpretations. This section performs the PC decomposition in Section 2.2.1 for each dataset and window in Table 2. This yields the coefficient series $(\beta_{t,k})_{t=1}^T$ and PCA basis $u_k(x), x \in \{-1.0, -0.99, \dots, 0.99, 1.0\}$ for $k = 1, \dots, K$.

We choose $K = 5$ as components to analyze, as this choice is sufficient for 90–95% of cumulative PVE for ARB5 and ETH5 over the windows considered. It also appears that for these windows, no reasonable choice of K will work for ETH30, but its analysis is considered a useful case study. We begin by analyzing the basis functions $u_k(x)$ to assess commonalities and interpretations. Subsection 4.4 analyzes the time series scores to attach a temporal meaning to the coefficient dynamics.

Figure 5 displays the first five basis functions for each dataset and window. As a baseline, ARB5 demonstrates a generally consistent shape and interpretation across windows:

- $u_1(x)$: Exhibits a *slope effect*, consistently tilting the liquidity surface. A shock in this factor increases liquidity toward $x = -1$ and simultaneously decreases it toward $x = +1$. This mode explains the largest proportion of variance.

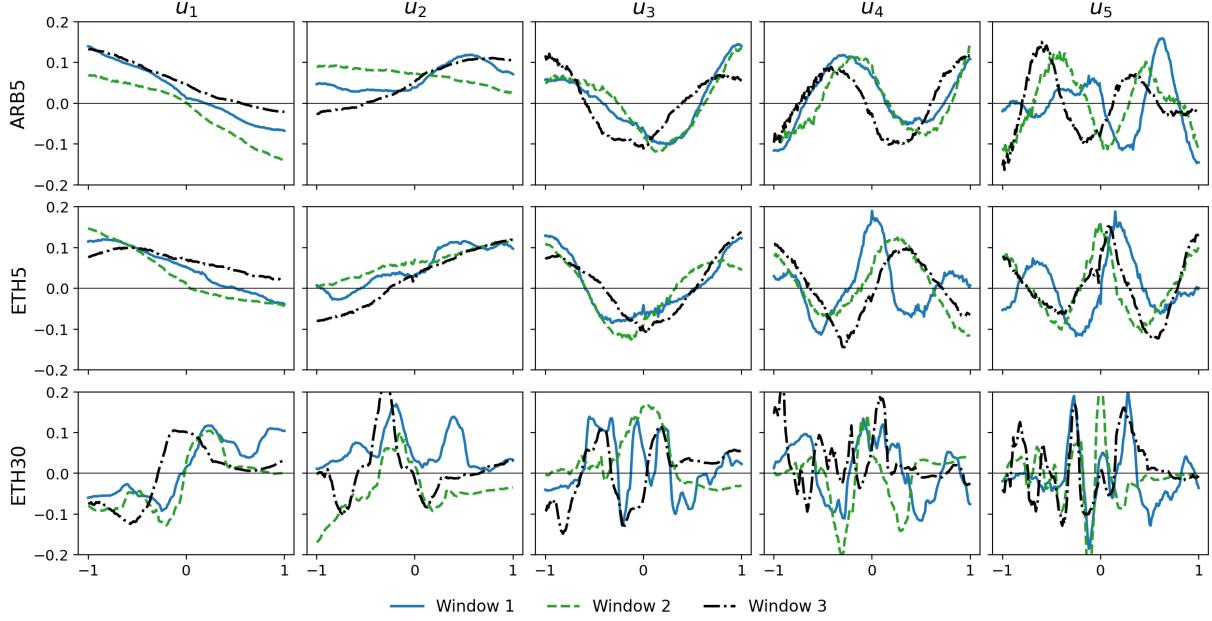


FIGURE 5. First five PCA basis functions for each dataset. Rows: dataset, columns: basis function index. Three windows are displayed per subpanel. Note that these are unique up to a sign change.

- $u_2(x)$: Represents a *mixed level/slope effect*: liquidity generally increases as these basis functions are strictly positive (aside from part of window 3). The asymmetry adds a slope, where one side is increased more than the other.
- $u_3(x)$: Characterized as a *curvature effect*: acts as a bell curve or symmetric beta distribution. A positive shock in this factor will decrease the liquidity toward the center and increase it at the edges.
- $u_4(x)$ and $u_5(x)$: Show increasing complexity with four and five inflection points, respectively, contributing to finer details of the liquidity surface.

Note that within each window, a more pronounced "level" effect can be obtained by a linear combination of u_1 and u_2 , offsetting their slopes. Comparing Window 3 with Windows 1 and 2 illustrates the idea of the same effect but in a rotated basis, where an approximate level effect can be found regardless of window, but each requires a different way to linearly combine the basis functions.

The ETH5 basis functions are also consistent across windows. The summary is highly similar to that of ARB5; we highlight a few differences:

- $u_1(x)$: Similar in shape to ARB5, but now mostly positive. A positive shock in this factor will increase liquidity in the $x = -1$ direction, with less of an effect toward the $+1$ direction.
- $u_2(x)$: Slightly more irregular than ARB5. Modes for Windows 1 and 2 are positive, whereas Window 3 crosses $y = 0$ closer to $x = 0$.

The remaining three basis functions are still similar, albeit with some slight irregularities and less discernable shapes. Across both ARB5 and ETH5, each of u_k , $k = 3, 4, 5$ tends to have $k - 2$ inflection points, consistent with a polynomial representation. One exception is

k	Window 1				Window 2				Window 3			
	PVE $_k$	ADF p	sd($\beta_{t,k}$)	τ	PVE $_k$	ADF p	sd($\beta_{t,k}$)	τ	PVE $_k$	ADF p	sd($\beta_{t,k}$)	τ
1	0.702	0.019	8.323	9.156	0.612	0.141	6.325	8.499	0.653	0.011	5.353	9.047
2	0.184	0.000	4.257	7.295	0.274	0.050	4.228	5.914	0.260	0.033	3.376	16.839
3	0.042	0.149	2.035	5.803	0.043	0.000	1.671	3.599	0.034	0.000	1.227	2.259
4	0.024	0.000	1.523	3.776	0.018	0.000	1.079	2.428	0.014	0.000	0.792	2.137
5	0.009	0.000	0.945	2.591	0.007	0.003	0.679	1.753	0.008	0.000	0.601	1.421

TABLE 3. Summary statistics for principal component time series of ARB5 for each window. τ is mean reversion time $-1/\log|\phi|$ from AR(1) fit (in units of 8 hours).

Window 1 for ETH5, which seems to have a higher polynomial order for its fourth and fifth basis functions, indicating a more complex functional shape during that period.

A consistent interpretation of ETH30 is challenging.

- $u_1(x)$: Often a logistic type shape. A positive shock in this factor would increase liquidity sharply for positive x and decrease it for negative, although the interpretation shifts slightly in Window 3 and in Window 2 appears to be making that transition.
- $u_2(x)$: In all windows, a positive shock would increase liquidity in a neighborhood of negative x close to $x = 0$. Otherwise, there are not many similarities across windows.

Higher order basis functions become rapidly uninterpretable, aside from factor 3 for window 2, which is surprisingly a nice mound shape. There seems to be a severe instability over time for ETH30's basis; this is further analyzed in Section 4.5.

4.4. PCA: Coefficient Time Series. For each principal component time series score associated with the factors above, we provide a time series analysis in two steps: (1) a lightweight overview for general observations (directly below), and (2) a second pass that targets specifics found in the first pass, including considering specific heteroskedastic models and innovation distributions (Section 4.4.3).

4.4.1. Arbitrum ETH-USDC 5bps (ARB5). Figure 6 presents the time series of the first five principal component scores ($\beta_{t,k}$) and their autocorrelations (ACFs) for the ARB5 dataset. Table 3 summarizes key statistics for the $\beta_{t,k}$, in particular, its proportion of variance explained, augmented Dickey-Fuller (ADF) test p -value, sample standard deviation $sd(\beta_{t,k})$ (over the window), and mean reversion time $\tau_k = -1/\log|\phi|$ when fitted to an AR(1) model. The ADF test assesses a null of a unit root against a trend stationary alternative. Note that we are not performing any formal hypothesis testing, although a larger p -value is suggestive of a unit root. For a benchmark comparison, we fix $p > 0.1$ as a cutoff to indicate potential unit root behavior (larger values moreso). These are always visually cross-checked with the time series and its ACF. Note that τ_k and $sd(\beta_{t,k})$ are in reference to an 8-hour period of time: $\tau_k/3$ is the daily reversion time. The standard deviation $sd(\beta_{t,k})$ generally does not scale with the sampling interval, so there is no straightforward rule to obtain a daily version.

There is a substantial jump (18–28%) in explained variance from the first to the second factor compared to the second to third (3–4%). The decay shape of the ACFs is also persistent

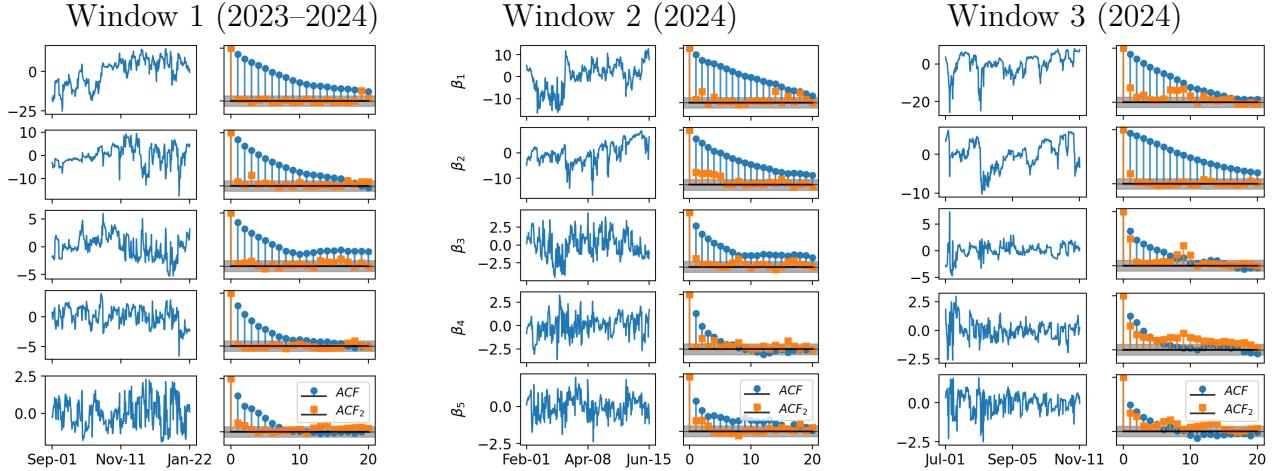


FIGURE 6. Time series (left) and autocorrelation functions of $\beta_{t,k}$ (right, blue) for ARB5. ACF_2 (orange) refers to the autocorrelation of the squared AR(1) residuals \hat{e}_t^2 . The gray bar indicates the 95% pointwise cutoff for significance from zero autocorrelation. Each ACF equals 1 at lag 0.

across windows: the fourth and fifth factor series typically exhibit rapid decay, while the first two are much slower. This is apparent in the table, with rapidly decaying mean reversion times, especially for $k = 3, 4, 5$. The standard deviations also show a similar decay at higher modes. Across windows, the standard deviations and mean reversion times are fairly similar, with some deviations per mode and window. The most consistent is the reversion time of the first factor series, with $\tau \approx 8.5\text{--}9.2$ (2.8–3 days).

The ADF p -values suggest that most series are stationary, with the exception of Factor 3 in Window 1 (figure: slow tapering of ACF and non-zero level is indicative of regime shift) and Factor 1 in Window 2 (linear decay in ACF like random walk).

The gray bar in the ACF plots represents the standard $1.96/\sqrt{n}$ asymptotic 95% pointwise cutoff for significant autocorrelation. The orange dots display the ACF of the squared residuals (\hat{e}_t^2) when the series is fitted to an AR(1). These suggest the presence of heteroskedasticity (GARCH effects) in nearly all series, some more pronounced than others.

4.4.2. First Pass Commonalities and Takeaways Across Datasets. There are many commonalities across datasets. As to not be repetitive with details, the analogous figures and tables for ETH5 and ETH30 are given in Appendix C (specifically, Figure 14 and Table 6 for ETH5, and 15 and 7 for ETH30), with a brief summary below.

ETH5 shares many similarities with ARB5, with some differences. Its fourth and fifth components have a significantly lower mean reversion time compared to the first three. For components 1 and 2, the times are mostly larger (varying from 0.75–3 times larger). The proportion of variances explained is also generally lower for the first two, and higher for the later components, although it is not extreme. The majority of coefficient series for ETH5 appear stationary, with one unit root in each window (component 1 in Windows 1 and 2, and component 2 in Window 3). The series standard deviations are similar in magnitude to ARB5.

ETH30 is less regular. There are now 1–3 nonstationary series in each window (evidence of unit root according to the ADF p), and visual inspection of the ACF verifies a slow linear decay in many of the series, some clearly nonstationary. The proportions of variance explained also increase in later components; for example, PVE₄, PVE₅ in Windows 1 and 2 increase by 2x compared to ETH5 (which already increased compared to ARB5), mirroring the discussion surrounding Figure 4.

A strong heteroskedastic effect is common across all three datasets, as evidenced by ACF₂ in the figures, which has several significant pieces (often at lower lags). The magnitude of ACF₂ across lags is remarkably similar across datasets and windows, appearing strongest in Window 3 and for the ARB5 dataset. Overall, several common features of the principal component time series emerge across all three datasets. The ACFs of the factor scores generally exhibit an exponential decay, characteristic of AR(1) processes. Additionally, significant autocorrelations in the squared residuals strongly suggest the presence of GARCH effects, indicating time-varying volatility in the principal components, a finding that is almost uniform across all components, datasets, and windows. For ARB5 and ETH5, most windows appear to have at most one series with a unit root, which is more pronounced for ETH5. In contrast, ETH30 consistently shows 1–3 series per window with this effect, and it is generally stronger. Finally, across all datasets, a notable shock in the first coefficient series is observed in Window 3 around August. From the start of Window 2, the first two coefficient series in both ETH5 and ETH30 share a common pattern of dropping from a neutral or positive value to a negative one, followed by a jump in the opposite direction in March. ARB5 shares a similar, though less pronounced, shift in its first coefficient series.

4.4.3. Detailed Time Series Models. To further investigate the statistical properties of the score series, we perform model selection across mean, volatility, and distribution specifications using the Bayesian Information Criterion (BIC) [KR95]. To ease notation, we drop the subscript k in $\beta_{t,k}$ until needed. For a given series $(\beta_t)_{t=1}^T$ and model \mathcal{M} (broadly containing distributional assumptions, parameters, etc.), write

$$\beta_t = \mu_t(\mathcal{F}_{t-1}; \mathcal{M}) + \varepsilon_t, \quad \varepsilon_t = \sigma_t(\mathcal{F}_{t-1}; \mathcal{M}) z_t, \quad z_t \stackrel{\text{iid}}{\sim} F(\mathcal{M}), \quad (9)$$

where $\mathcal{F}_u = (\beta_s, \sigma_s, \varepsilon_s)_{s \leq u}$. The GARCH(1,1) model yields the volatility recursion

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \gamma \sigma_{t-1}^2, \quad \omega > 0, \alpha, \gamma \geq 0. \quad (10)$$

The condition $\alpha + \gamma < 1$ yields a stationary covariance. The ARCH(1) model is simply a GARCH(1,0), meaning it drops the σ_{t-1}^2 term.

For a given time series model \mathcal{M} with d estimated parameters, let $\hat{\ell}$ be the maximized log-likelihood with sample size T . The Bayesian information criterion (BIC) is defined as

$$\text{BIC} = -2\hat{\ell} + d \log T. \quad (11)$$

BIC balances goodness of fit $-2\hat{\ell}$ with a complexity penalty $d \log T$. When comparing models on the same dataset, the model with lower BIC is preferred. More specifically, if two models \mathcal{M}_1 and \mathcal{M}_2 have equal prior probabilities, the posterior odds admit the approximation

[KR95]

$$\frac{\mathbb{P}(\mathcal{M}_1 \mid (\beta_t)_{t=1}^T)}{\mathbb{P}(\mathcal{M}_2 \mid (\beta_t)_{t=1}^T)} \approx \exp\left(\frac{1}{2}(\text{BIC}_2 - \text{BIC}_1)\right).$$

By monotonicity, one often relates these odds with the difference $\Delta_{1,2}\text{BIC} = \text{BIC}_2 - \text{BIC}_1$ as a measure of preference toward \mathcal{M}_1 (a lower BIC_1 gives larger $\Delta_{1,2}\text{BIC}$, increasing its odds). Kass and Raftery [KR95] give four evidence cutoffs, for $\Delta_{1,2}\text{BIC}$ which are $(0, 2)$, $[2, 6)$, $[6, 10)$, and $[10, \infty)$, corresponding respectively to labels “not worth a bare mention”, “positive”, “strong”, and “very strong”. Thus, to compare across many $\mathcal{M}_j, j = 1, \dots, J$ for the same $(\beta_t)_{t=1}^T$, it is sufficient to compare each model relative to that with the smallest BIC.

The subsequent analysis considers two model selection procedures using BIC as a comparative metric. The following is performed for all factor score series within each window and dataset:

- (1) Using an AR(1) mean (no trend or intercept), fit a collection of heteroskedastic models and distributions. This is done first because the current analysis strongly suggests an AR(1) style component with apparent heteroskedasticity but imprecise details.
- (2) Using volatility models based on step (1), double-check if the mean specification is correct.

For the first step, the idea is to check some overarching properties and commonalities across datasets, windows, and factors. Since there are no specific hypotheses to be tested, we sweep over a large collection of volatility specifications and innovation distributions that can assess a variety of properties. All modelling is done with the `arch` package [She21] in Python and utilizes the majority of available volatility processes: constant volatility, ARCH(1), GARCH(1,1), EGARCH(1,0,1), EGARCH(1,1,1), GJR-GARCH(1,1,1), and TARCH(1,1,1). See Appendix E for full details of the seven models tested. In summary, this collection of volatility models can identify certain traits based on the data’s model preference:

- not heteroskedastic (homoskedastic): constant volatility;
- no volatility clustering: ARCH(1);
- volatility clustering: all GARCH variants;
- asymmetry effects: EGARCH(1,1,1), GJR-GARCH(1,1,1), TARCH(1,1,1);
- multiplicative volatility dynamics: EGARCH (both 1,0,1 and 1,1,1);
- outlier sensitivity: TARCH is less sensitive than (GJR-)GARCH to extreme values; EGARCH further compresses through logarithm.

We also consider all of the available default distributions for the driver of the innovations, z_t . These are the normal, t , skew- t , and generalized error distribution (GED). Heavy tails are a hallmark of the t -distribution. For the GED, we restrict its shape parameter to be greater than 2 as an assessment for platykurtic behavior (t will check for leptokurtic). The skew- t informs whether there is an instantaneous asymmetry effect (compare with the volatility asymmetry, which is state-dependent).

Parameters are estimated by MLE using `arch` (Python). Appendix C shows a heatmap comparing all $3 \cdot 3 \cdot 5 = 45$ series across all $4 \cdot 6 = 24$ combinations of volatility and distributions.

The results are overwhelmingly consistent in favor of some GARCH effect with heavy-tailed innovations (t -distribution). More specifically:

- (1) GARCH or GARCH variants always (100%) produced the lowest BIC, when compared to ARCH(1) or a constant volatility.
- (2) The t or skew- t distribution was always (100%) preferred to the normal or platykurtic GED.
- (3) The top three choices were EGARCH(1,0,1)- t ($26/45 = 57.8\%$), EGARCH(1,1,1)- t ($6/45 = 13.3\%$), and TARCH(1,1,1)-skew t ($5/45 = 11.1\%$).
- (4) There were 11 cases where there was strong or very strong evidence against EGARCH(1,0,1)- t ($\Delta\text{BIC} > 6$). Except for one case with GARCH(1,1)- t , they were either EGARCH(1,1,1)- t or a skew- t paired with GARCH(1,1), EGARCH(1,0,1), or TARCH(1,1,1). Out of those 11 cases, 7 of them occurred in ARB5.

Aside from the last bullet, there were no clear patterns where preference was for or against a certain model according to dataset or time window.

In general, conditional heteroskedasticity with heavy tails is unequivocal. The common winner is EGARCH(1,0,1)- t , suggesting symmetric volatility with multiplicative growth and heavy tails. When asymmetry is preferred, it is mixed between the innovations (skew- t) or volatility process (TARCH(1,1,1) or EGARCH(1,1,1)) or both. The results also suggest a preference toward regularizing large shocks, since EGARCH generally wins, and TARCH is always preferred over GJR-GARCH.

To check if AR(1) is appropriate, we compare mean models using EGARCH(1,0,1) volatility with t -distributed noise based on the last step. This simply checks different AR(p) orders, specifically $p = 1, 2, 3$, as well as an ARMA(1, 1) model for a potential moving average effect. Note that $p = 3$ could have contextual relevance as three lag units equate to one day (so it could be related to a regional trading time, like the operational time differences of America, Europe, Asia). A BIC heatmap comparing choices is provided in Appendix C. Toward Occam's razor, consider the cases where there was a BIC difference of > 6 against AR(1): This occurred in $8/45 = 17.8\%$ of the series considered, and ARMA(1,1) was never preferred. An AR(3) model was only preferred in ETH5 Window 1, for $k = 2, 3, 4$. The remaining 5 cases of AR(2) were scattered across datasets, windows, and component number. Thus, AR(1) is a reasonable choice according to BIC, and is further supported by the ACF decay rates in Figure 6.

4.5. Basis Stability. With many commonalities existing both in the statistical properties of the scores and the PCA basis functions, we move toward assessing subspace stability and working with a common basis. We use the same rolling window approach as in Section 4.2, now keeping track of the full sequence of $T \times K$ matrices $(U_t)_{t \geq 0}$ over each rolling window start time t , where U_t is the matrix whose columns are the orthogonal eigenfunctions in the PCA decomposition as computed over the length T window $[t, t + T]$. We measure basis stability using the projection distance in Equation (16). To specifically measure how much a subspace is drifting, we use two reference bases:

- $d(U_t, U_0)$, the subspace distance between the basis U_t from PCA data over $[t, t + T]$ and U_0 , the basis estimated at inception $[0, T]$, and
- $d(U_t, U_{\mathcal{L}})$, where \mathcal{L} are the Legendre polynomials.

To motivate the latter, recall that Figure 5 showed similarities with polynomials of increasing degree, relating to terms of its inflection points. Additionally, rotating the first two basis functions together formed a type of “intercept plus slope.” The Legendre polynomials $\{P_n\}_{n=0}^{\infty}$ form an orthogonal basis over $[-1, 1]$ (when equipped with the Lebesgue measure), where P_n is of degree n . The first five are displayed in Figure 7. As intended, they share many similarities when compared with the empirical bases in Figure 5.

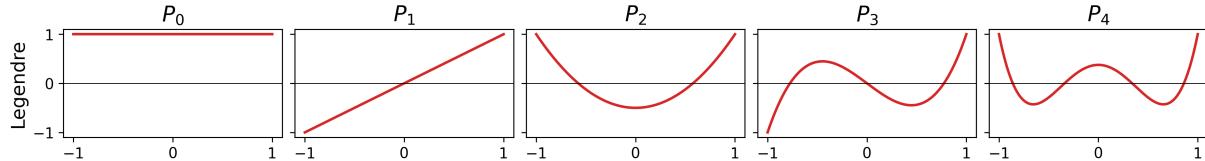


FIGURE 7. First five Legendre polynomials: $P_0(x) = 1$, $P_1(x) = x$, $P_2(x) = \frac{1}{2}(3x^2 - 1)$, $P_3(x) = \frac{1}{2}(5x^3 - 3x)$, $P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$.

Figure 8 displays the subspace distance (16) for $K = 3, 4, 5, 6, 7$ over the rolling window start date t . The Legendre coefficients were determined as in Equation (14) using Simpson’s rule. The top panel shows $d(U_t, U_0)$, the amount by which the basis has drifted since the initial eigenbasis was obtained at $t = 0$, and the bottom panel shows $d(U_t, U_L)$, comparing the empirical eigenbasis against the fixed Legendre basis.

For a reference benchmark, recall from Section A that $0 \leq d \leq K < M - K$, and the expected distance between two random K -dimensional subspaces is $K(1 - K/M)$, or 2.955, 3.920, 4.876, 5.821, 6.756 for $K = 3, 4, 5, 6, 7$ respectively.

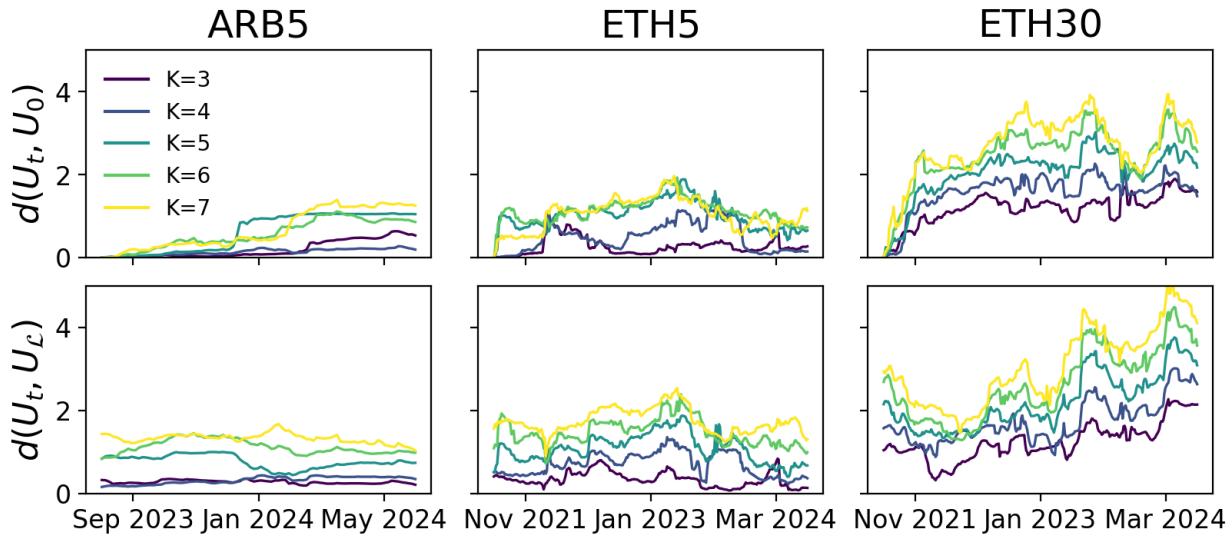


FIGURE 8. Distance between subspaces. U_t refers to the eigenmode basis performing PCA over times $\{t, t+1, \dots, t+T-1\}$, and U_L refers to the Legendre basis. x -axis values are equally spaced, corresponding to the rolling window start time. K refers to the dimension of the basis used.

Looking at the figure in order of datasets, ARB5 generally indicates high stability. All K exhibit relatively small $d(U_t, U_0)$ throughout. Notably, $K = 4$ is more stable than $K = 3$ relative to inception, suggesting information from higher modes is retained even far into the

future. A possible breakpoint appears in higher dimensions around Jan 2024, with a jump for $K = 5$ that later propagates to $K = 6$ and $K = 7$. Alignment with the Legendre basis is tight, as $d(U_t, U_{\mathcal{L}})$ remains low across time and K , similar in magnitude to $d(U_t, U_0)$ for later t . The fourth-dimensional subspaces are consistently of low distance, and even up to the seventh dimension appear quite stable.

ETH5 is qualitatively similar, but is generally noisier with more fluctuations. Interestingly, $d(U_t, U_0)$ shows transient drift that later reverts, a property hallmark of processes that are weakly stationary with slow mean reversion. ETH5 is also generally close to $U_{\mathcal{L}}$, with a similar drop in similarity toward middle dates. Interestingly, all three-dimensional subspaces for ETH5 are highly close even during periods of higher drift.

ETH30's nonstationarity is fully present in Figure 8. The subspace distances increase rapidly a few months from inception. Each distance is slowly drifting higher, potentially stabilizing to a value close to the random subspace expected distance (e.g. 2.955 for $K = 3$ and 6.756 for $K = 7$).

Appendix F repeats the rolling subspace distances in Figure 8 considering different rolling window sizes ($T = 200, 800$ compared to $T = 400$) and approximate subsampling intervals (4 and 16 hours compared to 8). The overall takeaways are the same regardless of these choices, with an increase in stability when $T = 800$ is used.

In summary, the bases for ARB5 and ETH5 are highly stable and comfortably align with the fixed Legendre basis up to at least the fourth dimension. This finding is significant because it provides a clear, interpretable, and natural starting point for modeling liquidity surfaces, a property that the ETH30 pool does not share. The ability to use a fixed, orthogonal basis like the Legendre polynomials simplifies analysis and enables consistent interpretation across different time periods. The next section illustrates its use and shows that its basis functions offer a natural interpretation of the liquidity surface.

4.6. Legendre Basis for Liquidity Surface. For ARB5 and ETH5, and each window, we fit to the Legendre basis and compare against the raw data. For indexing consistency, enumerate $k = 1, 2, \dots$ and write $\psi_k(x) = P_{k-1}(x)$ as the Legendre polynomial of degree $k - 1$. Specifically, this means to consider the rank K reconstruction

$$\hat{y}_t^{(K)}(x) = \sum_{k=1}^K \beta_{t,k} \psi_k(x), \quad y_t(x) = \hat{y}^{(K)}(x) + r_t^{(K)}(x), \quad x \in [-1, 1], \quad (12)$$

where the $\beta_{t,k}$ are determined as in Equation (14) using Simpson's quadrature rule, and $r_t^{(K)}$ is the remainder. Since non-zero intercepts in the $\beta_{t,k}$ provide useful insight, no demeaning is performed prior to the fit. Figure 9 shows the Legendre fit for ARB5 and ETH5, separately fitted over datasets and windows for the first cross-section in each window (analogous to the cross-sections in Figure 2). This is done for $K = 5$ (blue curve) and $K = 50$ (green), with the observed $y_t(x_m)$ overlayed (grey points).

The $K = 50$ reconstruction cleanly interpolates except in cases of large jumps or potential outlier spikes (particularly evident in the ETH5 case), providing a low bias but high variance fit. With $K = 5$, it is a smoother, low-variance fit. Visually, it is highly accurate for ARB5 and captures ETH5's coarse shape. Note that neither is objectively better; larger K lowers

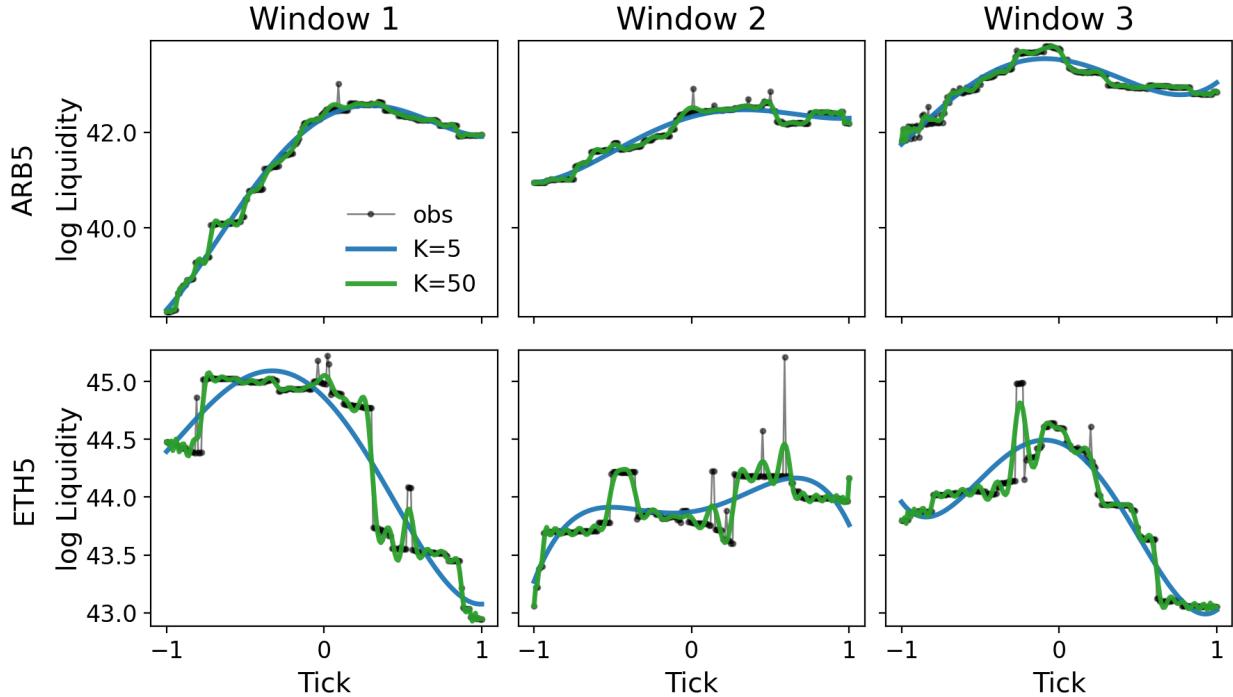


FIGURE 9. Cross-sections comparing Legendre reconstructions using $K = 5$ versus $K = 50$ over all windows and ARB5 and ETH5. Each cross-section uses the first t in that window.

approximation error but risks fitting noise; $K = M$ interpolates in-sample, a typical bias-variance tradeoff. The remainder of the text focuses on the first five components ($k = 1, \dots, 5$) since (1) orthogonality means taking $K > 5$ does not change their interpretation or coefficients, (2) empirically, $K = 5$ explains $\approx 90\text{--}95\%$ of variance across windows under the PCA eigenbasis, and the Legendre basis is closely aligned.

4.6.1. Factor Specifics and Effects of Shocks. In contrast with PCA eigenfunctions, the Legendre basis yields a fixed collection of functions across time. One may interpret the k th basis function in the following way:

- $\psi_1(x) = 1$ (*level*): A positive shock raises liquidity uniformly over prices, representing liquidity being added everywhere.
- $\psi_2(x) = x$ (*slope/tilt*): Increases liquidity toward $x = -1$ while decreasing toward $x = +1$ (or vice versa), matching the empirical “slope effect.”
- $\psi_3(x) = \frac{1}{2}(3x^2 - 1)$ (*curvature*): A positive shock would lift the edges toward $x = \pm 1$ and depress at the current price $x = 0$, the same as “flattening” the liquidity surface, which has concentration at zero. A shock in the opposite direction enforces a bell shape.
- $\psi_4(x) = \frac{1}{2}(5x^3 - 3x)$ (*inhomogeneous tilt*): This has the same tilt effect as $\psi_2(x)$ at $x = 0$ and ± 1 , but a reversed effect toward the two inflection points $x = \pm 1/\sqrt{5} \approx \pm 0.4472$. A positive shock leaves $x = 0$ untouched, increases liquidity near $x = -1/\sqrt{5}, 1$, and decreases liquidity near $x = 1/\sqrt{5}, -1$.

- $\psi_5(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$ (*mid-level repositioning*). Increasing the corresponding weight shifts liquidity symmetrically on $x \approx \pm\sqrt{3/7}$ toward the middle and edges. Effectively, liquidity near (but not at) the middle would shift to either the far edges or toward $x = 0$.

The interpretations become unwieldy as K increases. Note that the lower order pieces hold interpretation even if $K > 5$ by orthogonality of the $\psi_k(x)$.

To better understand these interpretations, Figure 10 works with ETH5 and reveals the effect of a positive “shock” to each factor, displaying the first cross-section in each window (mirroring the Figure 2). Specifically, $\beta_{t,k}$ is replaced with $\beta_{t,k} + \text{sd}(\beta_{t,k})$ in the reconstruction Equation (12). This is done separately for each k to observe its individual effect on the curve. The sample standard deviation is used for simplicity to illustrate the effect of a “typical” shock. The figure shows the observed $y_{t_0}(x_m)$ (grey points) as well as the Legendre projected versions $\hat{y}_t^{(5)}(x)$ (blue curve) and its shocked version (red curve). The ARB5 version is Figure 18 in Appendix D, showing the same patterns.

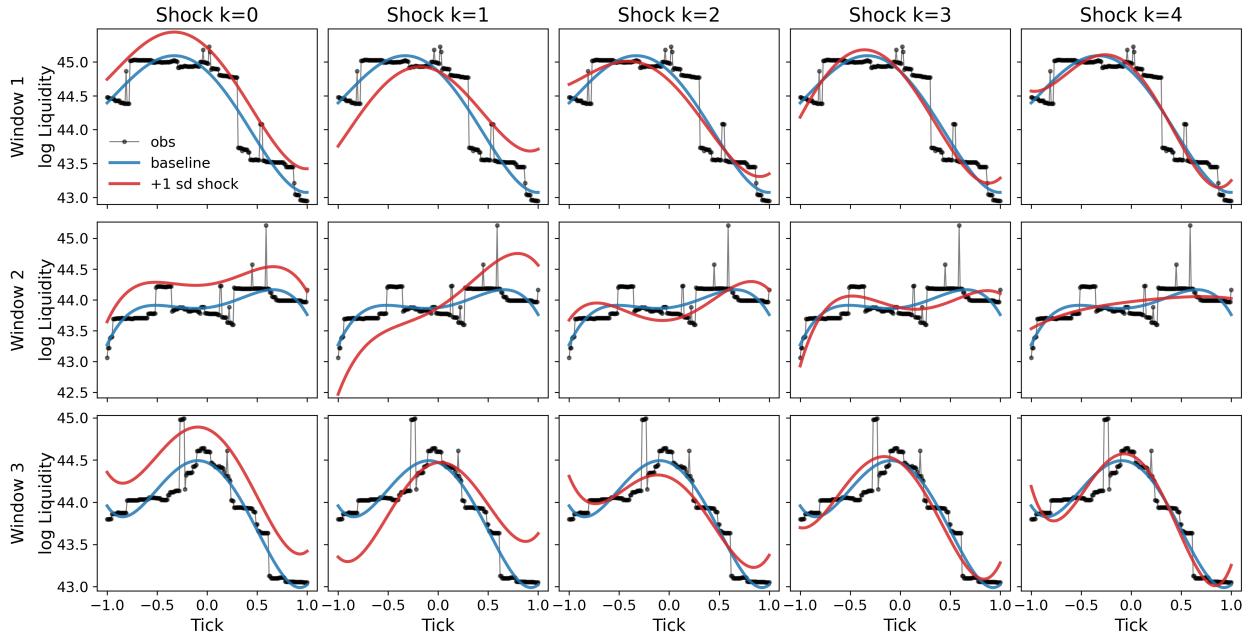


FIGURE 10. ETH5: effect onto the full cross-section $y_t(x)$ of a 1 standard deviation shock on the k th component, $k = 1, \dots, 5$, where t is the beginning of the window considered.

The observations match the bulleted discussion above, as expected. Notably, a shock of $k = 1$ increases the liquidity curve uniformly, $k = 2$ tilts it, and $k = 3$ distributes mass away from the current price at $x = 0$. The higher order effects are less obvious (likely because of a smaller standard deviation), although a close inspection reveals $k = 4$ providing the inhomogeneous tilt. This is especially apparent in Window 2 where $x = 0$ is left untouched, $x < 0$ shows a decrease at $x = -1$ and increase at $x = -1/\sqrt{5}$, and the reverse holds for $x > 0$. For $k = 5$, liquidity increases at both the far tails and center, redistributing mass near $x = \pm\sqrt{3/7}$ (local minima of $\psi_5(x)$).

4.6.2. *Legendre Time Series.* We briefly repeat the time series analysis as done in Section 4.4.

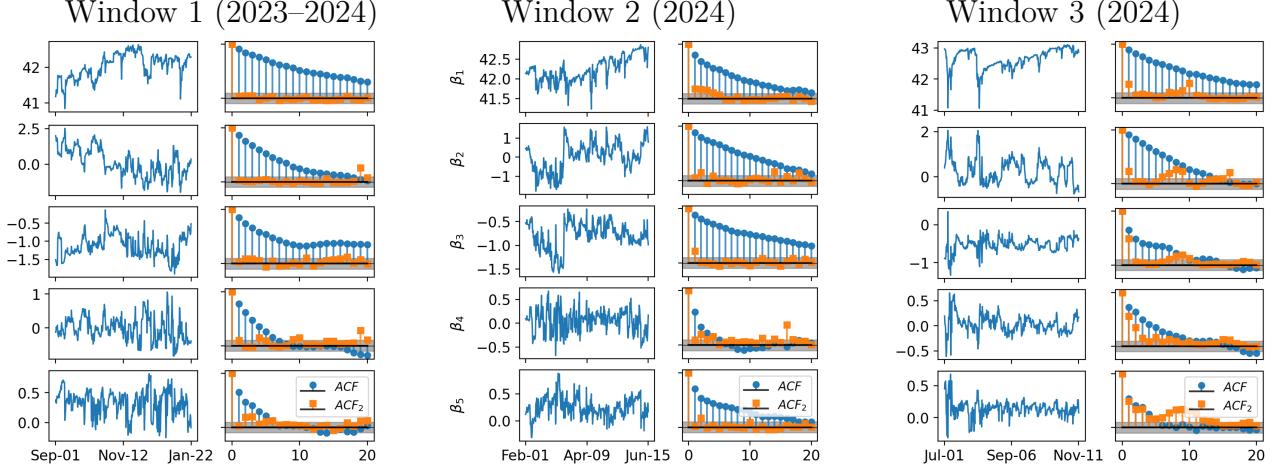


FIGURE 11. Legendre Fit: Time series (left) and autocorrelation functions of Legendre coefficients $\beta_{t,k}$ (right, blue) for ARB5. ACF₂ (orange) refers to the autocorrelation of the squared AR(1) residuals \hat{e}_t^2 . The gray bar indicates the 95% pointwise cutoff for significance from zero autocorrelation. Each ACF equals 1 at lag 0.

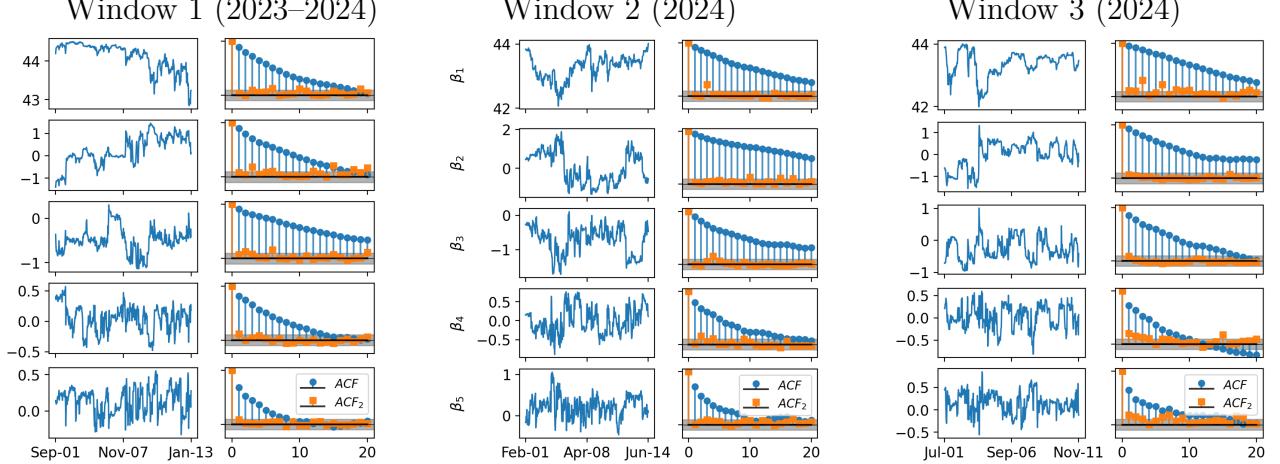


FIGURE 12. Legendre Fit: Time series (left) and autocorrelation functions of Legendre coefficients $\beta_{t,k}$ (right, blue) for ETH5. ACF₂ (orange) refers to the autocorrelation of the squared AR(1) residuals \hat{e}_t^2 . The gray bar indicates the 95% pointwise cutoff for significance from zero autocorrelation. Each ACF equals 1 at lag 0.

Many of the takeaways in these figures coincide with conclusions from the PCA discussion. We list some notable remarks:

- The same spikes in midsummer 2024 are present in Window 3, for both ARB5 and ETH5.

k	Window 1				Window 2				Window 3			
	ADF p	$\bar{\beta}_{t,k}$	$sd(\beta_{t,k})$	τ	ADF p	$\bar{\beta}_{t,k}$	$sd(\beta_{t,k})$	τ	ADF p	$\bar{\beta}_{t,k}$	$sd(\beta_{t,k})$	τ
1	0.030	42.044	0.346	11.437	0.225	42.235	0.311	4.555	0.022	42.651	0.315	9.051
2	0.009	0.094	0.948	7.387	0.054	0.039	0.743	8.399	0.002	0.249	0.544	11.378
3	0.075	-1.125	0.308	6.678	0.001	-0.793	0.281	7.831	0.000	-0.522	0.185	2.350
4	0.000	-0.028	0.347	3.890	0.000	0.049	0.231	2.034	0.000	0.036	0.174	3.115
5	0.000	0.327	0.210	2.390	0.000	0.250	0.183	2.908	0.000	0.145	0.125	1.603

TABLE 4. Legendre Fit: Summary statistics for Legendre coefficients of ARB5 for each window. τ is mean return time $-1/\log|\phi|$ from AR(1).

k	Window 1				Window 2				Window 3			
	ADF p	$\bar{\beta}_{t,k}$	$sd(\beta_{t,k})$	τ	ADF p	$\bar{\beta}_{t,k}$	$sd(\beta_{t,k})$	τ	ADF p	$\bar{\beta}_{t,k}$	$sd(\beta_{t,k})$	τ
1	0.652	44.128	0.349	9.569	0.118	43.278	0.372	16.322	0.023	43.362	0.397	23.758
2	0.069	0.235	0.634	11.443	0.033	0.026	0.804	17.784	0.184	-0.013	0.607	14.919
3	0.031	-0.499	0.274	10.886	0.000	-0.722	0.402	9.717	0.000	-0.257	0.350	6.550
4	0.000	0.042	0.210	5.029	0.001	0.001	0.343	4.205	0.000	0.031	0.260	3.125
5	0.000	0.135	0.176	2.609	0.000	0.214	0.261	3.012	0.000	0.146	0.230	2.357

TABLE 5. Legendre fit: Summary statistics for Legendre coefficients time series of ETH5 for each window. τ is mean return time $-1/\log|\phi|$ from AR(1).

- The time series and ACFs carry many of the same properties as the PCA version. A particular example is $k = 3$, ETH5, Window 1, where the PCA and Legendre scores are nearly indistinguishable.
- Heteroskedasticity is the biggest thing that seems to be adjusted, sometimes amplified or shifted across k (see e.g. ARB5 Window 3)

The tables remove the proportion of variance explained, which are not directly applicable to a fixed basis expansion, and add $\bar{\beta}_{t,k}$, the average over that period.

- The first coefficient is an intercept. The sample mean $\bar{\beta}_{t,1}$ is large and dominates the other $\bar{\beta}_{t,k}$, appropriate as it represents the level effect of liquidity.
- The average $\bar{\beta}_{t,3}$ for the curve effect is uniformly negative across windows and datasets and large relative to its standard deviation. This makes sense because increasing the curve effect flattens it out. A negative baseline is needed to obtain the bell curve shape.
- Conversely, the average mid-level repositioning $\bar{\beta}_{t,5}$ is uniformly positive. This counteracts the pure “quadratic” shape from the negative $\bar{\beta}_{t,2}$, creating a tail effect akin to a bell curve. It also has the fastest mean reversion time (approximately 1 day, sometimes faster for ARB5), suggesting a transient effect.
- The slope effects ($k = 2, 4$) always have large $sd(\beta_{t,k})$ relative to their averages. This suggests an insignificant average slope effect, i.e. the liquidity surface can drift from asymmetry but will revert toward it. The mean reversion time for $k = 2$ is quite high with $\tau \approx 7.5\text{--}18$. In contrast, the cubic $k = 4$ has $\tau \approx 2\text{--}5$.

- Mean reversion times decrease rapidly as k increases, particularly for $k = 4$ and 5.
- We still observe 0–1 unit roots across the first five coefficient series within each window/dataset. The ones with ADF $p > 0.1$ are the level effect ($k = 1$), common across ARB5 Window 2, and ETH5 Windows 1 and 2, and the slope effect ($k = 2$) for ETH5 Window 3. These look like a random walk/unit root (slow decay). For Window 3, it looks like a regime shift by the ACF (quick decay then level off at a nonzero value), and the switch in the time series.

Detailed Time Series Analysis. We repeat the investigation of the time series as in Section 4.4.3 for the Legendre time series. Rather than looking for broad patterns across all datasets and windows, the focus is shifted toward each basis function as they carry a fixed meaning in this context.

Figure 13 shows a heatmap comparing all $3 \cdot 2 \cdot 5 = 30$ series across all $4 \cdot 6 = 24$ combinations of volatility and distributions. In aggregate, the patterns are very similar to the PCA case, favoring a GARCH effect with heavy tails. The t or skew- t distribution are always preferred to the normal or platykurtic GED, and the choices still include EGARCH(1,0,1)- t ($12/30 = 40\%$), TARCH(1,1,1)-skew t ($6/30 = 20\%$), with GARCH(1,1)- t the third most preferred now ($4/30 = 13.3\%$). Interestingly, there are now two cases for an ARCH(1)- t , but both have a $\Delta BIC < 1$ compared to the EGARCH(1,0,1)- t . With fixed factor interpretations across windows, it is important to point out some specific observations:

- (1) For $\beta_{1,t}$, the coefficients associated with the level effect $P_0(x) = 1$, there is unanimous preference for the asymmetric TARCH(1,1,1) or EGARCH(1,1,1) over EGARCH(1,0,1) (with 5/6 preferring skew- t and the other a symmetric- t). All of these have $\Delta BIC > 6$ compared to the best EGARCH(1,0,1) in that row, aside from ARB5 Window 1 ($\Delta BIC = 5.1$). Thus, the level effect exhibits high asymmetry, both in the volatility (state dependent) and innovations (instantaneous shocks).
- (2) EGARCH(1,0,1) or GARCH(1,1) is always preferred for $\beta_{4,t}$ and $\beta_{5,t}$ over asymmetric EGARCH(1,1,1) or TARCH(1,1,1). The skew- t wins for ARB5, and t for ETH5. Thus, these high-order effects prefer symmetric volatilities, with potential skew in the innovations.
- (3) $\beta_{2,t}$ and $\beta_{3,t}$ are split across all choices mentioned above.

We repeat the mean analysis, using EGARCH(1,0,1) volatility with t -distributed noise. A detailed BIC heatmap is in Figure 19 in Appendix D. The takeaways are similar to those of the PCA discussion, with a general preference toward AR(1) with a couple of exceptions. This time, the first two factors ($k = 1, 2$) are consistently AR(1) (10/12 give lowest BIC, and the other two have $\Delta BIC = 1.8, 4.7$). The higher-order factors are more mixed, but still with a strong preference toward AR(1). AR(3) was never preferred.

5. DISCUSSION AND CONCLUSION

This research explored the application of *dynamic factor models* and *functional principal component analysis (FPCA)* to model and analyze liquidity profiles in Uniswap v3. Our findings demonstrate the potential of these methodologies to effectively capture the complex dynamics in liquidity distributions across price ranges. By treating liquidity profiles as

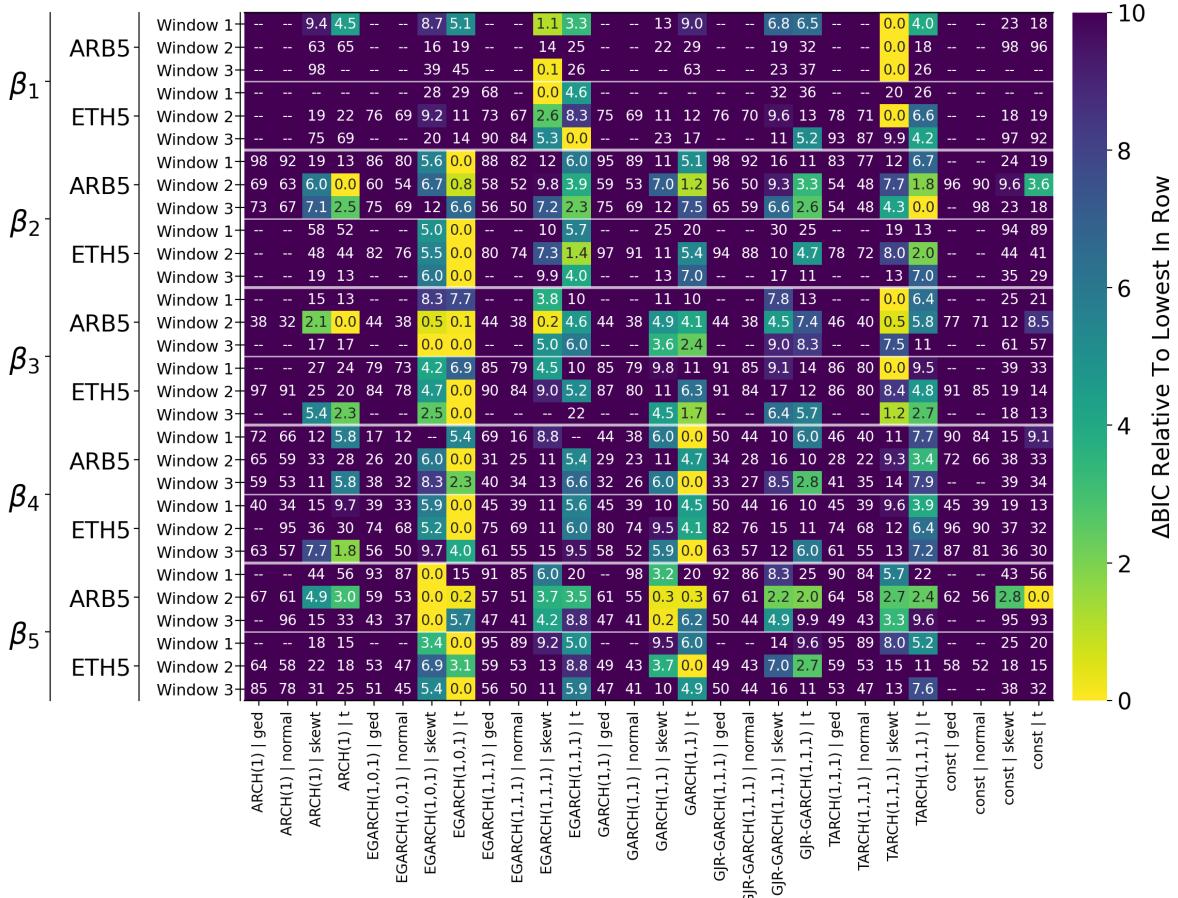


FIGURE 13. Legendre Coefficients: Heatmap of BIC scores according to heteroskedasticity and distribution assumptions. Reported is the Δ BIC relative to the lowest (best) for that series. All use an AR(1) mean. Any value of “-” had $\Delta BIC > 10$.

functional data, we can leverage tools for dimensionality reduction and forecasting that maintain the structure of the original data.

Our empirical analysis highlights a *persistent low-rank factor representation* within the Uniswap v3 liquidity surfaces of the 5 bps pools. For all datasets and time periods considered, the factor time series coefficients exhibited *autoregressive (AR) dynamics*, clear *generalized autoregressive conditional heteroskedasticity (GARCH) effects*, and *heavy tails*. This was consistent for both the empirical (PCA) bases and a Legendre basis. A statistical verification using the Bayesian Information Criterion (BIC) overwhelmingly favored GARCH variants over constant volatility models, confirming a preliminary observational analysis. The models also consistently preferred heavy-tailed innovation distributions, such as the Student's t -distribution, over the normal distribution, suggesting that liquidity surfaces are prone to occasional extreme shocks. Additionally, there appears to be approximately one unit root present in the collection of factor time series for the 5 bps pools, a commonality in other functional time series in finance, such as yield curves. In summary, heavy tails, conditional heteroskedasticity, and AR(1)-style mean-reversion are unequivocal features of (log)-liquidity surfaces.

A key finding is the *stability of the empirical PCA basis* for the 5 bps pools (Ethereum ETH-USDC and Arbitrum ARB-USDC). This low-rank structure is highly stable across rolling time windows and also aligns consistently with a low-order Legendre polynomial basis. This alignment is a critical discovery, as it provides a fixed, interpretable set of factors, akin to the level, slope, and curvature factors found in the Nelson-Siegel model for yield curves, which can be used for consistent analysis and modeling. This finding suggests that the fundamental "shape" of liquidity surfaces for these pools can be parsimoniously represented by a collection of orthogonal factors with persistent distributional properties. This sets the stage for accurate forecasting, uncertainty quantification, and analysis of economic shocks.

In contrast, the Ethereum 30 bps ETH-USDC pool revealed a highly nonstationary structure with significant temporal irregularities. Its PCA basis drifted substantially over time, requiring a higher number of principal components to explain the same proportion of variance observed in the 5 bps pools. This stark difference suggests that the higher 30 bps fee tier may attract liquidity providers with different patterns, resulting in a less stable and less consistent liquidity surface.

5.1. Stylized Facts of Uniswap v3 Liquidity Surfaces. The Legendre factor model can be used to empirically verify the stylized facts in the seminal work [FMCM⁺21, FMCA⁺22]. The statistical analysis of the log-liquidity surface confirms several stylized facts from existing literature and reveals new ones.

5.1.1. Verification of Known Stylized Facts.

- We confirm the claim that “*we typically observe $L_t(x)$ to be relatively large near $x = 0$.*” The average coefficients $\bar{\beta}_{t_k}$ for the asymmetric terms ($k = 2, 4$) in Tables 4 and 5 are both effectively zero across all time windows, meaning asymmetry averages out over time. We also uniformly observe that $\bar{\beta}_{t_3} < 0$ (significantly so) and $\bar{\beta}_{t_5} > 0$, both of which contribute to a larger liquidity mass around $x = 0$.
- We provide empirical evidence for the stylized fact that “*The shape of the liquidity surface can vary significantly depending on LP risk preferences and chosen price interval configurations.*” and that “*Market volatility significantly influences LP behavior*”. Our findings show that in high-volatility regimes, LPs tend to widen their liquidity ranges, shifting liquidity mass outwards and lowering the central peak. Conversely, in low-volatility periods, liquidity is characterized by a tighter, steeper concentration around the mid-price. According to our model, this phenomenon is captured by the “flattening effect” or $\psi_3(x)$ with coefficients $\beta_{t,3}$. A higher average $\beta_{t,3}$ value corresponds to a flatter liquidity surface, providing clear evidence for risk-averse LP behavior.

5.1.2. New Stylized Facts. Our statistical analysis reveals a succinct set of new facts about the liquidity surface dynamics.

- **Low-Rank Structure:** Dynamic factor analysis shows that 3 to 7 orthogonal components are sufficient to explain 90% to 95% or more of the variation in the log-liquidity surface. This percentage occasionally drops to 80% during periods of high volatility.

- **Persistent Time Series:** The time series associated with these factors exhibit AR(1)-like behavior when stationary. There is overwhelming evidence of GARCH(1,1) volatility and heavy-tailed innovations, and these facts hold consistently across the pools studied.
- **Basis Alignment:** For the 5 bps pools (ARB-USDC and ETH-USDC), the empirical PCA bases strongly align with the basis of Legendre polynomials. This consistency across time windows allows for the use of a parsimonious, fixed-factor model. The distributional facts mentioned above also hold for the Legendre coefficient series.
- **Nonstationarity:** Approximately one unit root is present in the collection of coefficient time series for the 5 bps pools, as found in both the PCA and Legendre bases. However, for the 30 bps ETH-USDC pool, the coefficient series are highly nonstationary, and there is no consistent basis, indicating a more complex and less stable liquidity structure.

5.2. Future Work. Our findings present several compelling avenues for future research, building upon the foundational framework we have established. An immediate next step is to examine whether the trends observed in the ETH-USDC 5bps and 30bps pools hold for other token pairs. While a full analysis of every possible pair is beyond the scope of this study, it would be valuable to empirically verify if less dominant pools are, in general, more nonstationary and less stable. We hypothesize that this is the case because the lower volume may attract liquidity providers who deploy more static or passive positions. This makes the liquidity dynamics more susceptible to price movements, which are themselves highly nonstationary.

From a similar perspective, it is worth investigating if the basis stability, particularly the alignment with the Legendre basis, holds for a wider array of dominant pools, such as those involving memecoins, stablecoin-stablecoin pairs, or the BTC-ETH pair. We anticipate that the stylized facts established in this study will remain consistent for high-volume pairs. More generally, extending the analysis to various token pairs and blockchain networks beyond Ethereum and Arbitrum would provide valuable insights into the generalizability of our findings and the potential impact of network-specific factors on liquidity dynamics.

From a modelling perspective, our work can be extended by exploring alternative orthogonal polynomial bases over $[-1, 1]$, such as the weighted Jacobi polynomials. These polynomials are orthonormal according to the weighted measure $w(x) = (1 - x)^\alpha(1 + x)^\beta$, where $\alpha, \beta > -1$. The standard Legendre basis is a special case where $\alpha = \beta = 0$. This offers additional flexibility, as the weights can be chosen to emphasize different parts of the liquidity curve. For example, setting $\alpha = \beta > 0$ places more weight toward the center of the curve, while $\alpha \neq \beta$ introduces an asymmetry, favoring one side of the price more than another. The optimal choice of basis ultimately depends on the practitioner's specific needs and the economic interpretation desired.

Moreover, the Jacobi polynomials are orthogonal eigenfunctions for the Jacobi differential operator:

$$\mathcal{L} = (1 - x^2) \frac{d^2}{dx^2} + (\beta - \alpha - \{\alpha + \beta + 2\}x) \frac{d}{dx}$$

associated with the eigenvalues $\lambda_n = -n(n + \alpha + \beta + 1)$ for n being nonnegative integers. It is thus tempting to use a parametric dynamical model for the evolution of the (log)

liquidity profile in the form of a stochastic partial differential equation (SPDE) driven by a finite-dimensional Brownian motion

$$dy_t(x) = \{\mathcal{L}y_t(x) - V(t, x)y_t(x) + h_t(x)\} dt + \sum_{r=1}^R \sigma_t^m(x)dW_t^r \quad \text{for } x \in [-1, 1],$$

where the differential operator \mathcal{L} depicts the local diffusion-convection behavior of the liquidity profile y_t , $V \geq 0$ corresponds to the percentage cancellation rate of liquidity at site x , and $h \geq 0$ captures the incoming flow rate of liquidity at x , and R is the number of Brownian drivers. The Jacobi polynomials, being the eigenfunctions for the Jacobi operator \mathcal{L} , help decompose the above SPDE into an infinite-dimensional system of SDEs in eigenmodes. As in the empirical analysis done in Section 4.6 using the Legendre bases and the fact that the eigenvalues λ_n diverge to negative infinity as $n \rightarrow \infty$, one would expect that the SDEs with the first few (approximately 5 to 7) eigenmodes will dominantly capture the dynamics of the (log) liquidity profile. The SPDE with the Jacobi differential operator thus provides a succinct and financially intuitive, easy-to-interpret model for the liquidity surface. We leave the details of this line of research for a future follow-up study.

Forecasting is a natural and crucial extension of this work, especially with the use of a fixed, interpretable basis. We propose a suitable forecasting recipe using a *vector autoregression (VAR) model* with a *multivariate GARCH process*. Consider the factor coefficient vector $\beta_t = [\beta_{t,1}, \dots, \beta_{t,K}]^\top$. An appropriate model is

$$\beta_{t+1} = \mathbf{a} + \mathbf{A}\beta_t + \epsilon_{t+1}, \quad \epsilon_{t+1} | \mathcal{F}_t \sim t_\nu(0, \mathbf{H}_{t+1}), \quad (13)$$

Here, \mathbf{a} is an intercept vector, \mathbf{A} contains the autoregression parameters, and \mathbf{H}_t follows a multivariate GARCH specification. This approach provides a full distributional forecast of the liquidity surface. This model would allow for:

- Full simulation of future paths for computing loss quantiles and conditional variance.
- Simulation of economic shocks by adjusting the current state or model parameters.
- Fast computation of look-ahead error metrics suitable for rolling forecast error calculations.

In this forecasting framework, the number of factors K would need to be determined beforehand, either by selecting a fixed, conservative number (e.g., $K = 5$) or based on a specific reconstruction metric. A thorough forecasting study using this methodology would be a valuable contribution to the field.

ACKNOWLEDGEMENTS

S. N. T. is grateful for the financial support from the National Science and Technology Council of Taiwan under grant 114-2115-M-007-012-MY3, "Mathematical Foundation of Automated Market Makers."

REFERENCES

- [AC20] Guillermo Angeris and Tarun Chitra. Improved price oracles: Constant function market makers. In *Proceedings of the 2nd ACM Conference on Advances in Financial Technologies*, AFT '20, pages 80–91, 2020.

- [ACDF25] Carol Alexander, Xi Chen, Jun Deng, and Qi Fu. Price discovery and efficiency in uniswap liquidity pools. *Journal of Futures Markets*, 2025.
- [AHPP20] Marco Avellaneda, Brian Healy, Andrew Papanicolaou, and George Papanicolaou. Pca for implied volatility surfaces. *arXiv preprint arXiv:2002.00085*, 2020.
- [AZR20] Hayden Adams, Noah Zinsmeister, and Dan Robinson. Uniswap v2 core. <https://uniswap.org/whitepaper-v2.pdf>, 2020. Accessed: July 17, 2025.
- [AZS⁺21] Hayden Adams, Noah Zinsmeister, Moody Salem, River Keefer, and Dan Robinson. Uniswap v3 core, 2021. Accessed: July 17, 2025.
- [BDG18] Jean-Philippe Bouchaud, Julius Bonart, Jonathan Donier, and Martin Gould. *Trades, Quotes and Prices: Financial Markets Under the Microscope*. Cambridge University Press, 2018.
- [BCN24] Erhan Bayraktar, Asaf Cohen, and April Nellis. Dex specs: A mean field approach to defi currency exchanges. *arXiv preprint arXiv:2404.09090*, 2024.
- [CDM24] Álvaro Cartea, Fayçal Drissi, and Marcello Monga. Decentralized finance and automated market making: Predictable loss and optimal liquidity provision. *SIAM Journal on Financial Mathematics*, 15(3):931–959, 2024.
- [CDR11] Jens H. E. Christensen, Francis X. Diebold, and Glenn D. Rudebusch. The affine arbitrage-free class of nelson–siegel term structure models. *Journal of Econometrics*, 164(1):4–20, 2011.
- [CF02] Rama Cont and José Da Fonseca. Dynamics of implied volatility surfaces. *Quantitative Finance*, 2(1):45, 2002.
- [CJ21] Agostino Capponi and Ruizhe Jia. The adoption of blockchain-based decentralized exchanges. *arXiv preprint arXiv:2103.08842*, 2021.
- [CM21] Rama Cont and Marvin S. Müller. A stochastic partial differential equation model for limit order book dynamics. *SIAM Journal on Financial Mathematics*, 12(2):744–787, 2021.
- [Con01] Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223, 2001.
- [CST10] Rama Cont, Sasha Stoikov, and Rishi Talreja. A stochastic model for order book dynamics. *Operations Research*, 58(3):549–563, 2010.
- [DL06] Francis X. Diebold and Canlin Li. Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130(2):337–364, 2006.
- [DLY08] Francis X. Diebold, Canlin Li, and Vivian Z. Yue. Global yield curve dynamics and interactions: a dynamic nelson–siegel approach. *Journal of Econometrics*, 146(2):351–363, 2008.
- [FMCA⁺22] Zhou Fan, Francisco J. Marmolejo-Cossío, Ben Altschuler, He Sun, Xintong Wang, and David Parkes. Differential liquidity provision in uniswap v3 and implications for contract design. In *Proceedings of the Third ACM International Conference on AI in Finance*, pages 9–17, 2022.
- [FMC⁺21] Zhou Fan, Francisco Marmolejo-Cossío, Daniel J. Moroz, Michael Neuder, Rithvik Rao, and David C. Parkes. Strategic liquidity provision in uniswap v3. *arXiv preprint arXiv:2106.12033*, 2021.
- [GM23] Emmanuel Gobet and Anastasia Melachrinos. Decentralized finance & blockchain technology. In *SIAM Financial Mathematics and Engineering 2023*, Jun 2023.
- [GPW⁺13] Martin D. Gould, Mason A. Porter, Stacy Williams, Mark McDonald, Daniel J. Fenn, and Sam D. Howison. Limit order books. *Quantitative Finance*, 13(11):1709–1742, 2013.
- [HKH15] Siegfried Hörmann, Łukasz Kidziński, and Marc Hallin. Dynamic functional principal components. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(2):319–348, 2015.
- [HL24] Hugo Hissinaga and Márcio Laurini. Interest rate forecasting with principal component analysis based on long-run covariance matrix. *Annals of Financial Economics*, 19(02):2450005, 2024.
- [HLR15] Weibing Huang, Charles-Albert Lehalle, and Mathieu Rosenbaum. Simulating and analyzing order book data: The queue-reactive model. *Journal of the American Statistical Association*, 110(509):107–122, 2015.
- [HRS⁺21] C. R. Harvey, A. Ramachandran, J. Santoro, F. Ehksam, and V. Buterin. *DeFi and the Future of Finance*. Wiley, 2021.
- [HSH12] Spencer Hays, Haipeng Shen, and Jianhua Z. Huang. Functional dynamic factor models with application to yield curve forecasting. *The Annals of Applied Statistics*, 6(3):870–894, 2012.

- [HSW22] Lioba Heimbach, Eric Schertenleib, and Roger Wattenhofer. Risks and returns of uniswap v3 liquidity providers. In *Proceedings of the 4th ACM Conference on Advances in Financial Technologies*, pages 89–101, 2022.
- [KGC⁺18] Harry Kalodner, Steven Goldfeder, Xiaoqi Chen, S Matthew Weinberg, and Edward W Felten. Arbitrum: Scalable, private smart contracts. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1353–1370, 2018.
- [KM23] Elie Kapengut and Bruce Mizrahi. An event study of the ethereum transition to proof-of-stake. *Commodities*, 2(2):96–110, 2023.
- [KR95] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.
- [LHRW21] Stefan Loesch, Nate Hindman, Mark B Richardson, and Nicholas Welch. Impermanent loss in uniswap v3. *arXiv preprint arXiv:2111.09192*, 2021.
- [Lit91] Robert Litterman. Common factors affecting bond returns. *Journal of Fixed Income*, pages 54–61, 1991.
- [MC23] Deborah Miori and Mihai Cucuringu. Defi: Modeling and forecasting trading volume on uniswap v3 liquidity pools. *Available at SSRN 4445351*, 2023.
- [MC24] Deborah Miori and Mihai Cucuringu. Clustering uniswap v3 traders from their activity on multiple liquidity pools, via novel graph embeddings. *Digital Finance*, 6(1):113–143, 2024.
- [Mec19] Elizabeth S Meckes. *The random matrix theory of the classical compact groups*, volume 218. Cambridge University Press, 2019.
- [MMR23] Jason Milionis, Ciamac C. Moallemi, and Tim Roughgarden. Complexity-approximation trade-offs in exchange mechanisms: Amms vs. lobs. In *Financial Cryptography and Data Security: 27th International Conference, FC 2023, Bol, Brač, Croatia, May 1–5, 2023, Revised Selected Papers, Part I*, pages 326–343. Springer-Verlag, 2023.
- [NS87] Charles R. Nelson and Andrew F. Siegel. Parsimonious modeling of yield curves. *Journal of Business*, pages 473–489, 1987.
- [Opr22] Andreea Oprea. The use of principal component analysis (pca) in building yield curve scenarios and identifying relative-value trading opportunities on the romanian government bond market. *Journal of Risk and Financial Management*, 15(6):247, 2022.
- [RS05] James O Ramsay and Bernard W Silverman. *Functional data analysis*. Springer, 2005.
- [She09] Haipeng Shen. On modeling and forecasting time series of smooth curves. *Technometrics*, 51(3):227–238, 2009.
- [She21] Kevin Sheppard. bashtage/arch: Release 4.18 (Version v4.18). Zenodo, March 2021.
- [SW11] James H Stock and Mark W Watson. Dynamic factor models. 2011.
- [TW24] Shen-Ning Tung and Tai-Ho Wang. A mathematical framework for modelling clmm dynamics in continuous time. *arXiv preprint arXiv:2412.18580*, 2024.
- [ZCY23] Haochen Zhang, Xi Chen, and Lin F Yang. Adaptive liquidity provision in uniswap v3 with deep reinforcement learning. *arXiv preprint arXiv:2309.10129*, 2023.
- [ZLW⁺24] Brian Z Zhu, Dingyue Liu, Xin Wan, Gordon Liao, Ciamac C Moallemi, and Brad Bachu. What drives liquidity on decentralized exchanges? evidence from the uniswap protocol. *arXiv preprint arXiv:2410.19107*, 2024.

APPENDIX A. BASIS DISCUSSION

Values at $x \notin \{x_1, \dots, x_M\}$ are not allowable in (6) without regularity: infinitely many L^2 functions interpolate the grid exactly but differ for $x \in (x_{m-1}, x_m)$. Thus, off-grid evaluation requires some smoothness or structure on u_k (or on y_t) and a numerical scheme.

One approach is to choose an orthogonal basis $\{\psi_k\}_{k \geq 1}$ of $L^2([-1, 1])$, for example Legendre, Fourier, or wavelets. The optimal L^2 projection coefficients are $\tilde{\beta}_{t,k} = \int_{-1}^1 y_t(x) \psi_k(x) dx$. Since this is intractable, as no off-grid evaluations are allowed, one typically approximates $\tilde{\beta}_{t,k} \approx \beta_{t,k}$ via a quadrature rule to determine weights $(w_m)_{m=1}^M$, e.g. trapezoidal or Simpson's

rule (both work in the equally spaced case). In particular,

$$\beta_{t,k} = \sum_{m=1}^M w_m y_t(x_m) \psi_k(x_m) \quad (14)$$

or for a fixed K in vector form

$$\boldsymbol{\beta}_t = (\Psi^\top W \Psi)^{-1} \Psi^\top W \mathbf{y}_t, \quad (15)$$

where Ψ is $M \times K$ with m, k entry $\psi_k(x_m)$, and $W = \text{diag}(w_1, \dots, w_M)$. This can be connected with the general FPCA approach, which assumes $y_t \in L^2([-1, 1])$ is stationary with mean m and covariance kernel $C : [-1, 1]^2 \rightarrow \mathbb{R}$, with C symmetric, positive definite, square-integrable, and continuous. Analogous to PCA, Mercer's Theorem yields $C(x, x') = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi_k(x')$ with orthonormal $\{\phi_k\}$. The Karhunen-Loève expansion subsequently provides $y_t(x) = m(x) + \sum_{k=1}^{\infty} \eta_{t,k} \phi_k(x)$, which is the L^2 -optimal basis when truncated to K terms. With equally spaced grids $\{x_m\}_{m=1}^M$, the eigenpairs of the sample covariance $\hat{\Sigma}$ converge (in operator norm) to those of the covariance operator as $M, T \rightarrow \infty$ (relative to the quadrature rule).

In summary, a fixed basis $\{\psi_k\}_{k \geq 1} \subset L^2([-1, 1])$ allows full evaluation of (6) at off-grid points. Although this is suboptimal compared to the Mercer basis, it still provides a consistent reconstruction in the limit, assuming sufficient regularity. Additionally, if the two bases align closely, one has the added benefit of a fixed interpretation if the $\psi_k(x)$ provide it.

Comparing Subspace Alignment. In practice, a data-dependent basis is not fixed over time due to sampling or estimation error, or nonstationarity. One may also want to use an alternative data-dependent basis, such as diagonalizing the *long-run* covariance (found to perform well in forecasting tasks [HL24], but suboptimal for low-rank representation of $y_t(x)$ in (4)), or a fixed basis $\{\psi_k\}_{k \geq 1}$ for a consistent interpretation, like Nelson-Siegel for yield curves, but still want it to be “close” to the empirical bases. Thus, it is useful to have a method to compare across bases.

Note that in any K -rank reconstruction (6), all $K \times K$ orthogonal rotations of factors yield the same reconstruction. Thus, when two bases reconstruct the same y , it means they span the same subspace. The degree to which these subspaces do not align can be compared through *projection distance*. Specifically, let $U_1, U_2 \in \mathbb{R}^{M \times K}$ be two rank- K matrices with orthogonal columns, each representing a particular basis, for example U_K from the PCA decomposition above, or Ψ from the fixed orthonormal basis. A standard metric to compare subspaces is the projection distance:

$$d_S(U_1, U_2) = \frac{1}{2} \|P_1 - P_2\|_F^2, \quad P_j = U_j U_j^\top \in \mathbb{R}^{M \times M}, \quad (16)$$

where $\|\cdot\|_F^2$ is the Frobenius norm. It follows that $0 \leq d_S(U_1, U_2) \leq \min(K, M - K)$. The subspaces are the same if and only if $d_S = 0$.

Thus, subspace drift across windows \mathcal{W}_1 and \mathcal{W}_2 can be measured by comparing $d_S(U_{\mathcal{W}_1}, U_{\mathcal{W}_2})$. As mentioned, one can also assess alignment with a fixed Φ . For example, if $d_S(U_{\mathcal{W}}, \Phi)$ is generally small over all windows \mathcal{W} , then the basis Φ is relatively stable for the data-generating process.

One benchmark stricter than the maximal distance $\min(K, M - K)$ is one according to a random subspace: if $U \in \mathbb{R}^{M \times K}$ corresponds to a random K -rank projection (Haar-distributed random matrix), then $E[d_S(U_j, U)] = K(1 - \frac{K}{M})$ (see Section 3.3 in Meckes (2019) [Mec19]). Values below this indicate more meaningful alignment than compared to a random subspace.

APPENDIX B. LIQUIDITY SURFACE DETAILS AND RANK-STANDARDIZED COORDINATES

To set notation, denote the *raw liquidity surface* as $\{\mathcal{L}_t(i)\}_{t \geq 0, i \in s\mathbb{Z}}$, where $\mathcal{L}_t(i)$ is the raw step function corresponding to liquidity at time t and tick i . Let s be the pool's tick-spacing (e.g. $s = 10$ for the 0.05% tier) and i_t the current price tick at time t . To simplify the modelling approach, we rank-standardize ticks, focusing on locations where liquidity changes across ticks.

Define the jump (event) ticks

$$E_t := \{i \in s\mathbb{Z} : \mathcal{L}_t(i) \neq \mathcal{L}_t(i - s)\},$$

i.e., the boundary ticks of the piecewise-constant \mathcal{L}_t (flat interiors excluded). Order jumps around \tilde{i}_t using a *rank* index $r \in \mathbb{Z}$:

$$\dots < i_t^{(-2)} < i_t^{(-1)} < i_t^{(0)} := \tilde{i}_t < i_t^{(1)} < i_t^{(2)} < \dots,$$

where, for $r > 0$ (resp. $r < 0$), $i_t^{(r)}$ is the $|r|$ -th jump strictly to the right (resp. left) of \tilde{i}_t .

Fix an odd $M \geq 3$ and let $j^* := (M + 1)/2$. We retain the anchor and the $(M - 1)/2$ nearest jumps on each side. For $j = 1, \dots, M$ define the orderings $i_{t,j} := i_t^{(j-j^*)}$ and the rank-standardized coordinates

$$x_j := \frac{2(j - j^*)}{M - 1} \in [-1, 1], \quad j = 1, \dots, M.$$

Note that $x_{j^*} = 0$ corresponds to the current price. By construction, $\{x_j\}_{j=1}^M$ is a fixed, equally spaced grid. Assume M is small enough that at each t there are at least $(M - 1)/2$ jumps on each side of \tilde{i}_t . The corresponding liquidity surface to be modelled over $x \in [-1, 1]$ is $L_t(x_j) := \mathcal{L}_t(i_{t,j})$.

This setup simplifies modelling but has two caveats: (i) to forecast future liquidity in absolute ticks, one must jointly model price (as in standard LOB settings), and (ii) it tracks jump locations (where the piecewise curve moves). In practice, jumps typically occur at each step near the current price, in which case x_j is simply an affine transform of the raw tick j .

APPENDIX C. ADDITIONAL TABLES AND FIGURES FOR ETH5 AND ETH30 PER WINDOW (PCA)

Figures 14 (ETH5) and 15 (ETH30) and Tables 6 (ETH5) and 7 are displayed here. The figures show time series, and ACF and ACF of squared residuals analogous to that of Figure 6 for ARB5. The tables show various summary statistics, analogous to Table 3 for ARB5. A summary of takeaways is in Section 4.4.

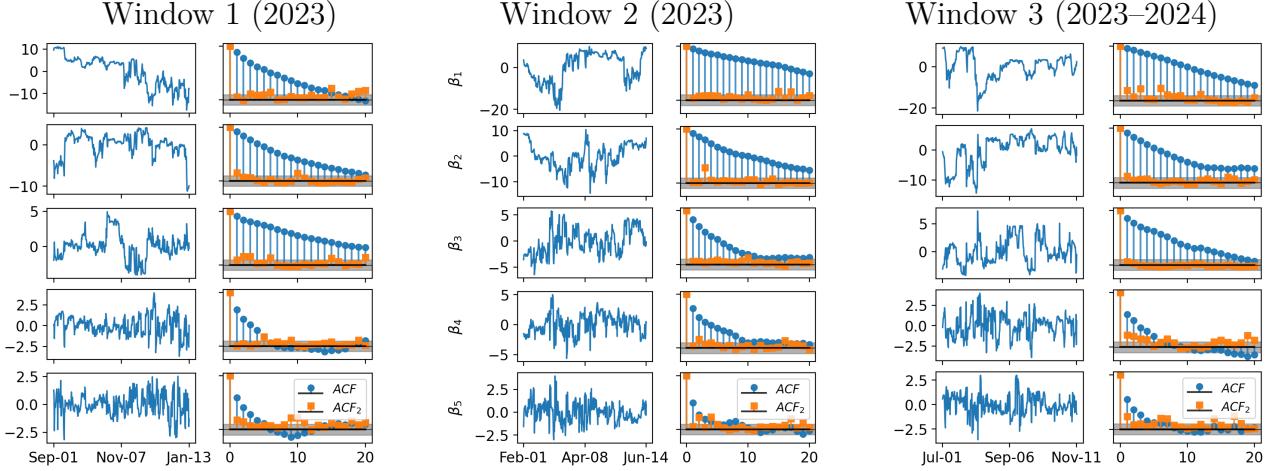


FIGURE 14. Time series (left) and autocorrelation functions of $\beta_{t,k}$ (right, blue) for ETH5. ACF_2 (orange) refers to the autocorrelation of the squared AR(1) residuals \hat{e}_t^2 . The gray bar indicates the 95% pointwise cutoff for significance from zero autocorrelation. Each ACF equals 1 at lag 0.

k	Window 1				Window 2				Window 3			
	PVE _k	ADF p	sd($\beta_{t,k}$)	τ	PVE _k	ADF p	sd($\beta_{t,k}$)	τ	PVE _k	ADF p	sd($\beta_{t,k}$)	τ
1	0.708	0.333	6.676	8.360	0.563	0.262	7.157	22.285	0.474	0.012	5.762	27.096
2	0.125	0.064	2.808	20.500	0.247	0.003	4.743	14.349	0.342	0.193	4.892	12.638
3	0.050	0.023	1.778	10.718	0.066	0.000	2.457	5.627	0.073	0.024	2.264	6.540
4	0.021	0.000	1.163	2.560	0.032	0.000	1.709	3.256	0.027	0.000	1.367	1.973
5	0.015	0.000	0.982	1.912	0.012	0.000	1.049	1.428	0.015	0.000	1.019	1.658

TABLE 6. Summary statistics for principal component time series of ETH5 for each window. τ is mean reversion time $-1/\log|\phi|$ from AR(1) fit.

k	Window 1				Window 2				Window 3			
	PVE _k	ADF p	sd($\beta_{t,k}$)	τ	PVE _k	ADF p	sd($\beta_{t,k}$)	τ	PVE _k	ADF p	sd($\beta_{t,k}$)	τ
1	0.582	0.385	3.068	17.055	0.430	0.130	6.396	29.026	0.591	0.358	7.878	123.980
2	0.166	0.431	1.636	64.624	0.150	0.259	3.783	36.975	0.120	0.001	3.553	12.187
3	0.049	0.000	0.890	9.819	0.110	0.100	3.239	20.934	0.087	0.009	3.030	15.286
4	0.038	0.049	0.787	11.956	0.064	0.013	2.469	13.638	0.025	0.014	1.623	7.152
5	0.030	0.000	0.702	5.480	0.040	0.000	1.948	5.334	0.020	0.000	1.461	5.360

TABLE 7. Summary statistics for principal component time series of ETH30 for each window. τ is mean reversion time $-1/\log|\phi|$ from AR(1) fit.

APPENDIX D. ADDITIONAL FIGURES FOR LEGENDRE DECOMPOSITION

Figure 18 gives the analogous version of 10 for ARB5. The shocks have the same effect according to the interpreted basis functions, but are generally less impactful because of lower variance.

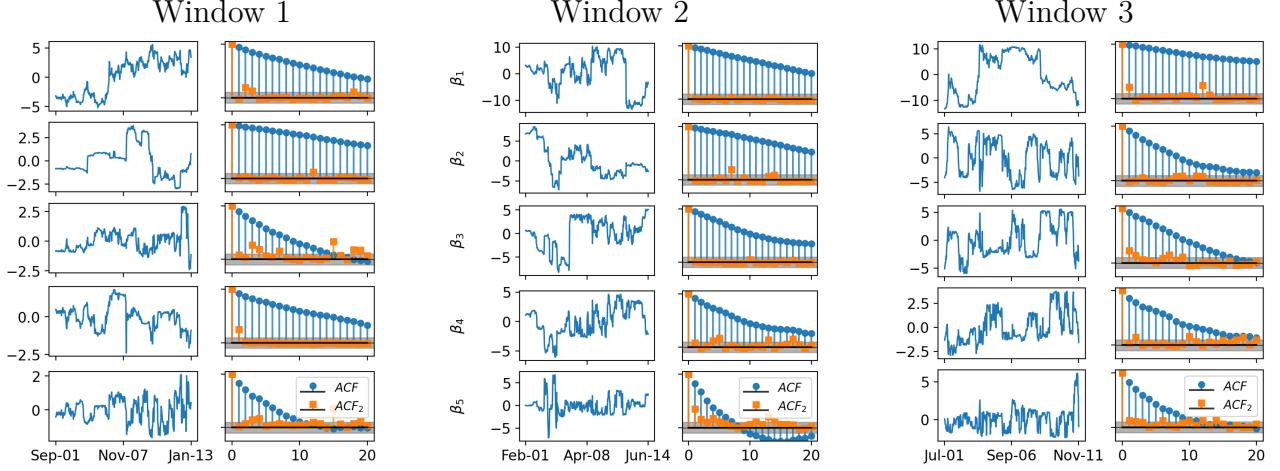


FIGURE 15. Time series (left) and autocorrelation functions of $\beta_{t,k}$ (right, blue) for ETH30. ACF_2 (orange) refers to the autocorrelation of the squared AR(1) residuals $\hat{\epsilon}_t^2$. The gray bar indicates the 95% pointwise cutoff for significance from zero autocorrelation. Each ACF equals 1 at lag 0.

Figure 19 gives a BIC heatmap according to mean choice with an EGARCH(1,0,1) volatility and t -distributed innovations. The takeaways are summarized at the end of Section 4.6.2.

APPENDIX E. TIME SERIES MODEL DETAILS

For each series $(\beta_{t,k})_{t=1}^T$ and model \mathcal{M} ,

$$\beta_{t,k} = \mu_t(\mathcal{F}_{t-1}; \mathcal{M}) + \varepsilon_{t,k}, \quad \varepsilon_{t,k} = \sigma_{t,k}(\mathcal{F}_{t-1}; \mathcal{M}) z_{t,k}, \quad z_{t,k} \stackrel{\text{iid}}{\sim} F(\mathcal{M}), \quad \mathbb{E} z_{t,k} = 0, \quad \text{var}(z_{t,k}) = 1.$$

An AR(1) without intercept satisfies

$$\mu_t = \phi \beta_{t-1,k}.$$

Define standardized shocks $e_{t,k} := \varepsilon_{t,k}/\sigma_{t,k}$.

E.1. Volatility Models. All constraints below are the usual ones ensuring positivity/stationarity. EGARCH does not need positivity constraints because of the log link.

ARCH(1).

$$\sigma_{t,k}^2 = \omega + \alpha \varepsilon_{t-1,k}^2, \quad \omega > 0, \quad \alpha \geq 0.$$

GARCH(1,1).

$$\sigma_{t,k}^2 = \omega + \alpha \varepsilon_{t-1,k}^2 + \beta \sigma_{t-1,k}^2, \quad \omega > 0, \quad \alpha, \beta \geq 0, \quad \alpha + \beta < 1 \quad (\text{sufficient for covariance stationarity}).$$

GJR-GARCH(1,1,1) (variance recursion with threshold).

$$\sigma_{t,k}^2 = \omega + \alpha \varepsilon_{t-1,k}^2 + \gamma \varepsilon_{t-1,k}^2 \mathbf{1}\{\varepsilon_{t-1,k} < 0\} + \beta \sigma_{t-1,k}^2,$$

with $\omega > 0, \alpha \geq 0, \alpha + \gamma \geq 0, \beta \geq 0$.

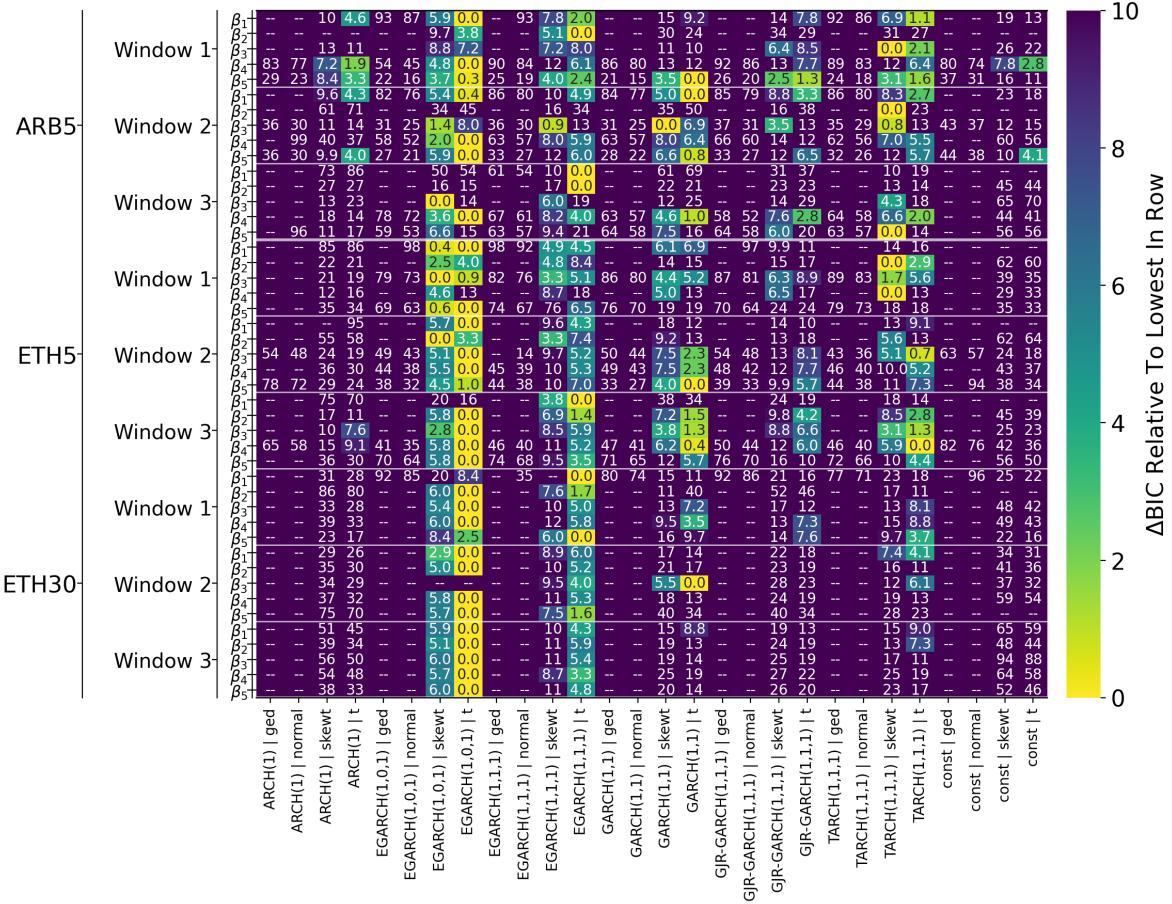


FIGURE 16. PCA Scores: Heatmap of BIC scores according to heteroskedasticity and distribution assumptions. Reported is the ΔBIC relative to the lowest (best) for that series. All use an AR(1) mean. Any value of “-” had $\Delta BIC > 10$.

TARCH(1,1,1) (standard-deviation recursion).

$$\sigma_{t,k} = \omega + \alpha |\varepsilon_{t-1,k}| + \gamma |\varepsilon_{t-1,k}| \mathbf{1}\{\varepsilon_{t-1,k} < 0\} + \beta \sigma_{t-1,k},$$

with $\omega > 0$, $\alpha \geq 0$, $\alpha + \gamma \geq 0$, $\beta \geq 0$.

EGARCH(1,0,1) (log-variance, symmetric).

$$\log \sigma_{t,k}^2 = \omega + \alpha(|e_{t-1,k}| - \kappa) + \beta \log \sigma_{t-1,k}^2,$$

where $\kappa = \mathbb{E}|Z|$ under the reference $Z \sim F$ (e.g., $\kappa = \sqrt{2/\pi}$ for normal).

EGARCH(1,1,1) (log-variance with leverage).

$$\log \sigma_{t,k}^2 = \omega + \alpha(|e_{t-1,k}| - \kappa) + \gamma e_{t-1,k} + \beta \log \sigma_{t-1,k}^2.$$

For GARCH(p, q) (and modifications depending on (p, o, q)): increasing p adds additional lagged shock terms (squared ε , absolute $|\varepsilon|$, or $|e|$ depending on the model), increasing o adds additional asymmetry/threshold lags, and increasing q adds additional volatility lags (variance, standard deviation, or log-variance accordingly).

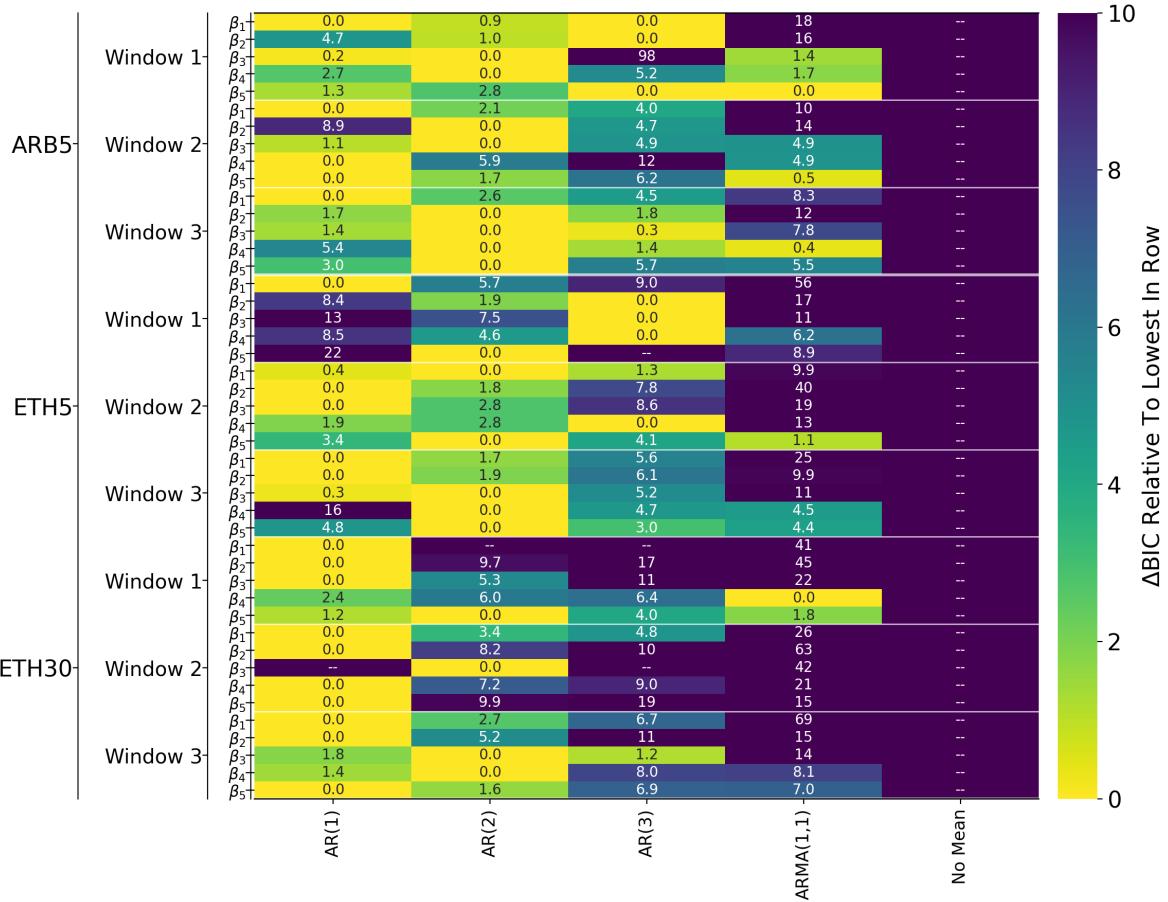


FIGURE 17. PCA Scores: Heatmap of BIC scores according to mean choice. Reported is the Δ BIC relative to the lowest (best) for that series. All use an EGARCH(1,0,1) volatility and t -distributed errors. Any value of “–” had $\Delta BIC > 10$.

Innovation distributions. All are standardized to unit variance in estimation:

- Normal;
- Student- t_ν ($\nu > 2$);
- Skew- t (Hansen) with shape (ν, λ) ;
- GED with shape κ_{GED} , here restricted to $\kappa_{GED} > 2$ to allow platykurtic checks.

APPENDIX F. ALTERNATIVE DATASET SIZES AND SUBSAMPLING

This section motivates some choices used in the paper. In particular, we consider:

- $T = 200$ and $T = 800$,
- an (approximate) subsampling frequency of 4 hours and 16 hours, and
- using $M = 101, 51, 11$ $x_m \in [-1, 1]$.

The main text uses $T = 400$, a subsampling frequency of 8 hours, and $M = 201$.

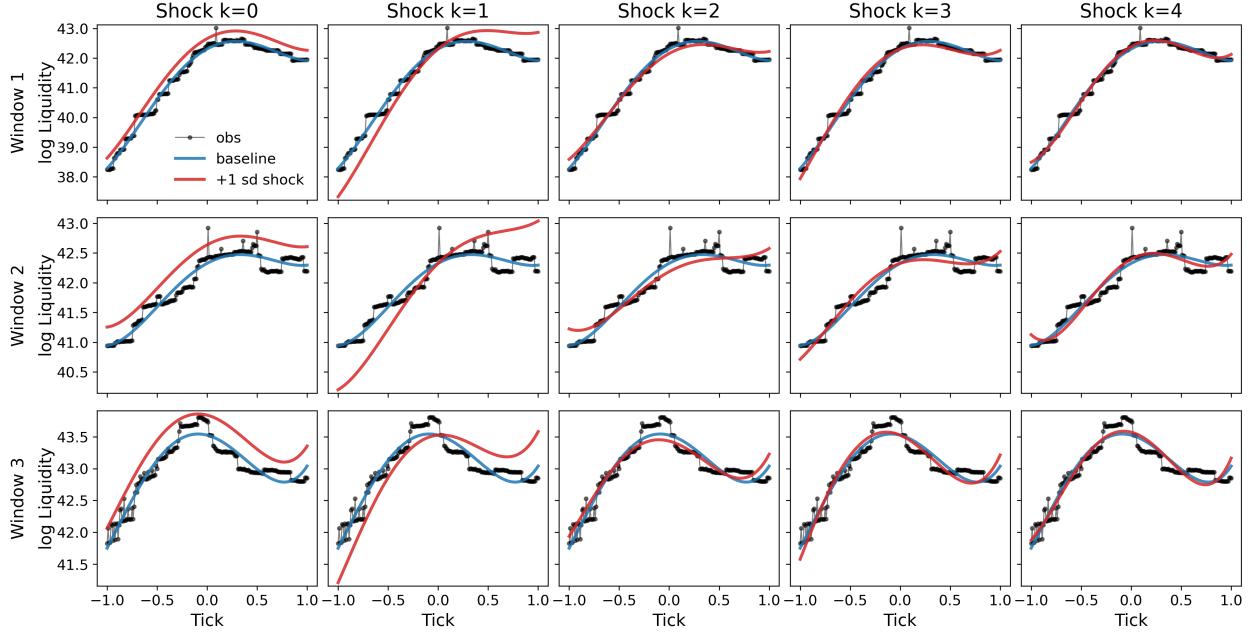


FIGURE 18. ARB5: effect onto the full cross-section $y_t(x)$ of a 1 standard deviation shock on the k th component, $k = 1, \dots, 5$, over the three windows considered.

Figure 20 illustrates PVE plots analogous to the bottom panel of Figure 4 for various choices of M , with $M = 201$ repeated for reference. The figure yields two major takeaways that hold across all datasets:

- PVE is generally higher in the lower M case, particularly $M = 11$.
- The change in PVE becomes much less discernible, where the $M = 101$ vs. $M = 201$ case is nearly indistinguishable

This makes sense theoretically, since in the discrete data regime the PCA tasks can be thought of as projecting the M -dimensional subspace onto a K dimensional one. For example, $M = 11$ means to summarize these 11 variables in terms of K linear combinations. When the data has some amount of correlation across x 's (as we expect it should), there should be little loss of information by truncating to, say, $K = 7$.

The second takeaway can be illustrated by example through comparing $M = 201$ and $M = 101$. In particular, the mild difference between the two is effectively saying that the $201 - 101 = 100$ prices further away from the center 101 prices are contributing little information. Again this is expected since the liquidity curves flatten out and revert to the center price. Even during periods of high volatility there is little difference between $M = 201$ and $M = 101$. Thus, $M = 201$ is sufficient for this task.

Figure 21 considers various choices of T and subsampling intervals, producing subspace distance plots analogous to Figure 8. The overall takeaways are the same as in the main body ($T = 400$, subsample frequency 8 hours), notably: (1) a period of drift that reverts for ETH5, (2) ETH30 drifting off, and (3) the Legendre basis producing similar results to the PCA basis. There are no clear differences between subsampling frequencies. For sample

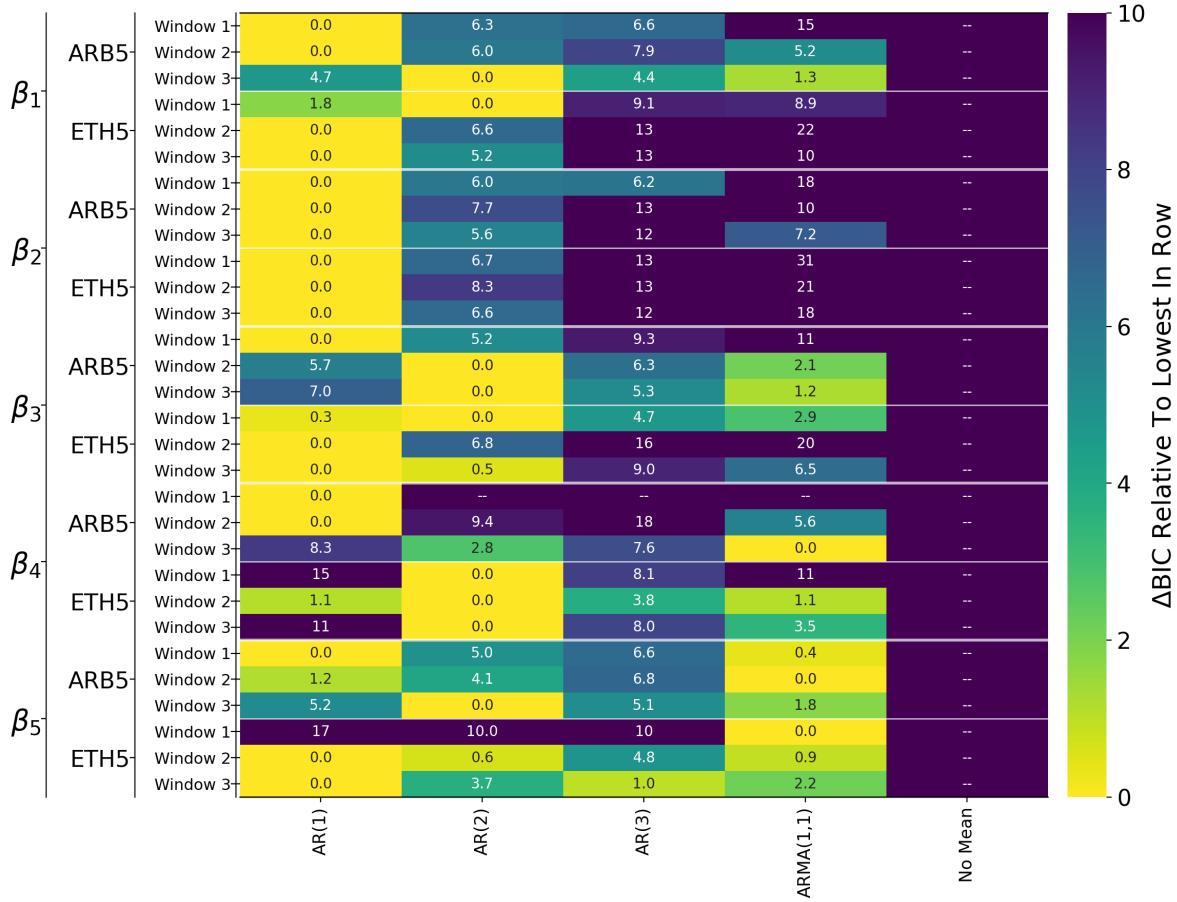


FIGURE 19. Legendre Coefficients: Heatmap of BIC scores according to mean choice. Reported is the Δ BIC relative to the lowest (best) for that series. All use an EGARCH(1,0,1) volatility and t -distributed errors. Any value of “–” has $\Delta BIC > 10$.

sizes, $T = 200$ is much rougher and $T = 800$ smoother, appearing to partially subdue some of the drift present in ETH30.

JIMMY RISK

DEPARTMENT OF MATHEMATICS AND STATISTICS,
CAL POLY POMONA
3801 W TEMPLE AVE, POMONA CA 91768

Email address: jrisk@cpp.edu

SHEN-NING TUNG

DEPARTMENT OF MATHEMATICS,
NATIONAL TSING HUA UNIVERSITY
HSINCHU, TAIWAN

Email address: tung@math.nthu.edu.tw

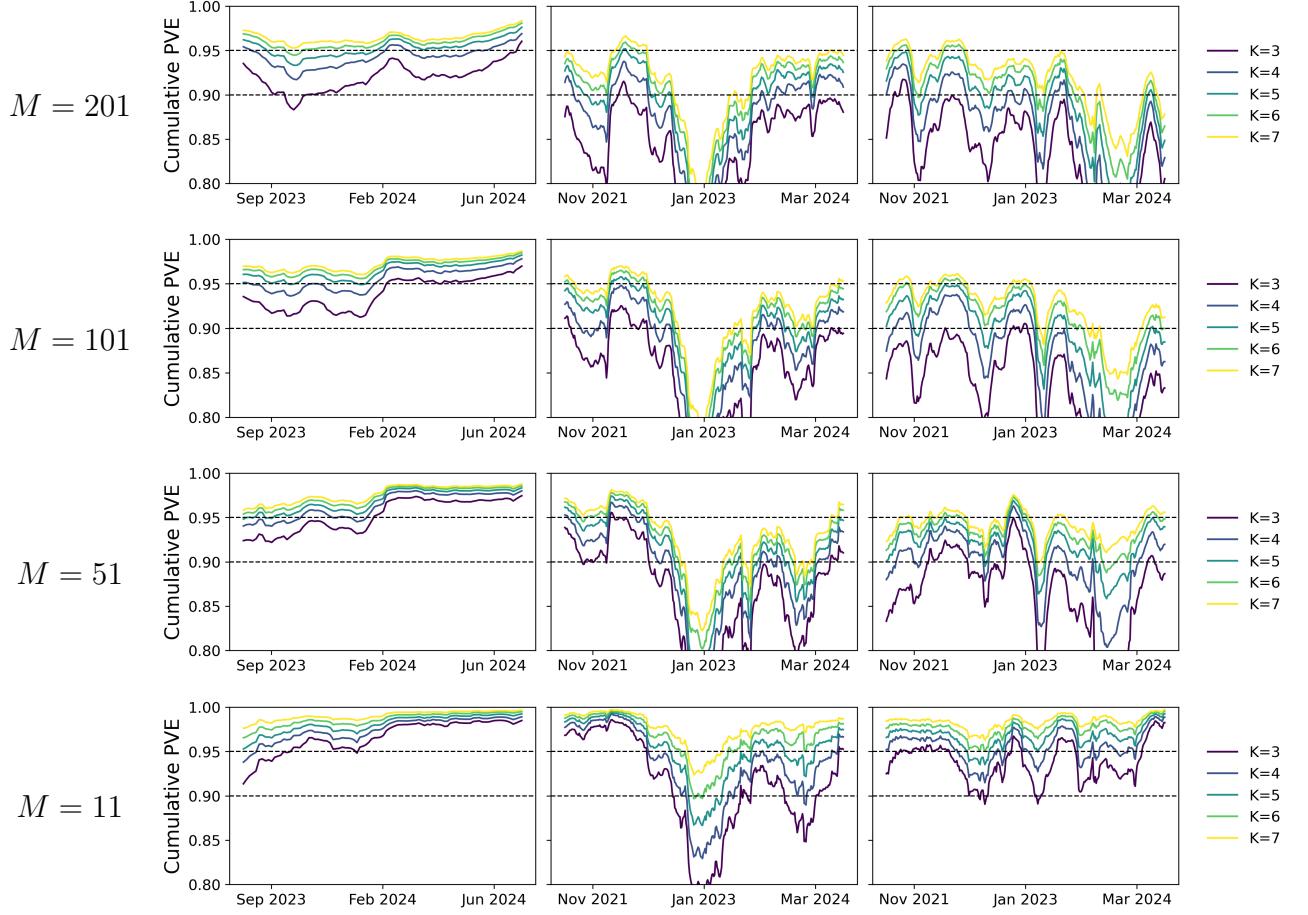


FIGURE 20. Left to right: ARB5, ETH5, ETH30. Proportion of variance explained over various choices of M (the number of x 's to use centered around the current price). Analogous to bottom panel of Figure 4.

TAI-HO WANG

DEPARTMENT OF MATHEMATICS
BARUCH COLLEGE, THE CITY UNIVERSITY OF NEW YORK
1 BERNARD BARUCH WAY, NEW YORK, NY10010

Email address: tai-ho.wang@baruch.cuny.edu

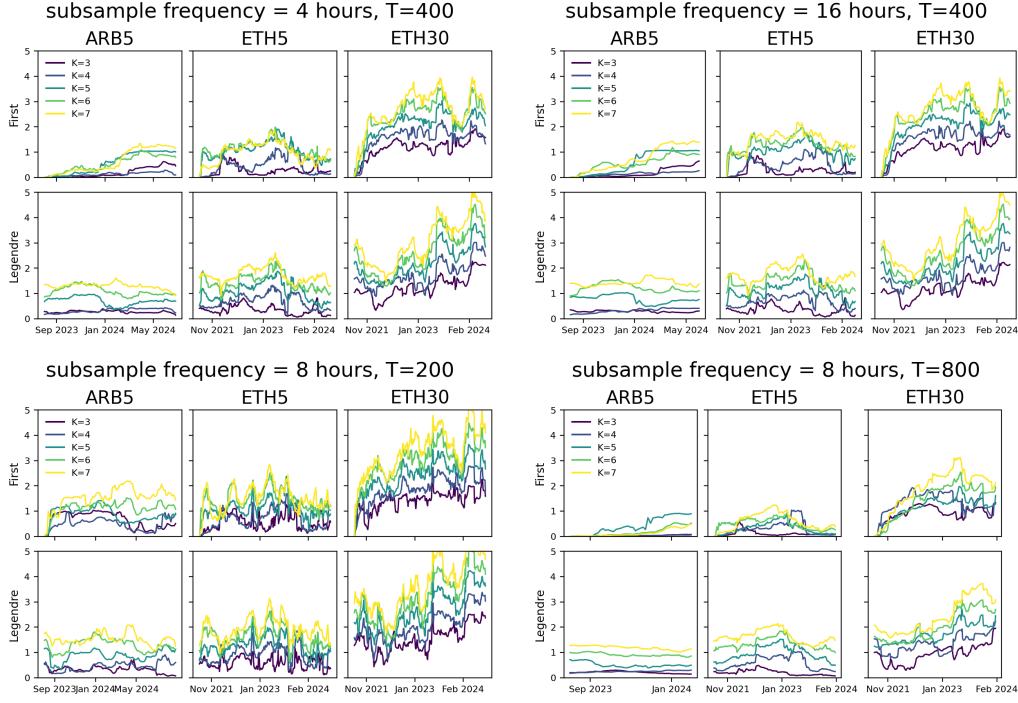


FIGURE 21. Subspace distances analogous to Figure 8 over various choices of subsampling frequency and rolling window size T .