

Three Numbers for Risk: A Data-Driven Spectral Basis for Portfolio Loss Distributions

Tamás Nagy, Ph.D.

tnagyphd@gmail.com

Working Paper

Abstract

The Spectral Fenton Distribution represents a portfolio loss distribution with 128 Fourier coefficients. Using simulated market data (5 assets, 250 trading days, stochastic volatility and correlation dynamics), we show that the time series of these coefficients is almost entirely one-dimensional: a single principal component captures 97.8% of all distributional variation, and three components capture 99.998%. This means the daily change in a portfolio’s risk profile is described by three numbers — $z_1(t)$, $z_2(t)$, $z_3(t)$ — instead of 128 coefficients, a $43\times$ compression with VaR reconstruction error below 0.05%. The result is robust: out-of-sample holdout error remains below 0.1%, and the 3-mode structure persists across portfolio sizes from 3 to 20 assets. The dominant mode z_1 correlates with VaR at $r = 0.998$, confirming it captures the overall risk level. We construct a real-time risk dashboard where each day’s entire distributional change is a point in \mathbb{R}^3 , and anomalous days (regime shifts) appear as outliers in this space. The learned basis is the natural state space for the Bayesian Live Risk, Dynamic URRT, and Schrödinger Bridge frameworks proposed in companion papers: once the basis is extracted, all of these become three-dimensional problems.

Key Messages

- 128 spectral coefficients compress to 3 numbers with $\leq 0.05\%$ VaR error (simulated data, $n = 5$ assets)
- Mode 1 alone captures 97.8% of all distributional variation; three modes capture 99.998%
- Out-of-sample holdout and robustness tests confirm stability across portfolio sizes ($n = 3$ to 20)
- The 3-number dashboard enables real-time regime detection via Mahalanobis distance
- The learned basis unifies Dynamic URRT, Bayesian Risk, and optimal transport into 3D problems

1. Introduction

1.1 The Dimensionality Question

The Spectral Fenton Distribution (Nagy, 2026a) encodes a portfolio’s loss distribution as $N = 128$ Fourier-cosine coefficients A_0, \dots, A_{127} . This is already a dramatic compression from the infinite-dimensional space of probability distributions. But for real-time risk management, even 128 numbers may be too many: a risk manager needs to understand, at a glance, what changed today.

The question is: **how many independent ways can the distribution actually change?**

If the answer is 3, then the risk dashboard is three numbers. If the answer is 10, it is ten. The answer determines the intrinsic dimensionality of the risk monitoring problem.

1.2 The PCA Approach

Let $A(t) \in \mathbb{R}^{128}$ denote the coefficient vector on day t . Over T trading days, we observe a $T \times 128$ matrix. Principal Component Analysis (PCA) extracts the dominant modes of variation:

$$A(t) \approx \bar{A} + z_1(t) \cdot v_1 + z_2(t) \cdot v_2 + \cdots + z_r(t) \cdot v_r$$

where \bar{A} is the time-averaged coefficient vector, v_1, \dots, v_r are orthonormal directions in \mathbb{R}^{128} (the principal modes), and $z_1(t), \dots, z_r(t)$ are the coordinates on day t . The question reduces to: how large must r be?

1.3 Related Work

PCA is a foundational tool in quantitative finance. The seminal application to interest rate term structures by Litterman and Scheinkman (1991) showed that three factors (level, slope, curvature) capture \$99% of yield curve variation; Dai and Singleton (2000) connected this to affine term structure theory. Alexander (2001) [TODO:cite Alexander, 2001] extended PCA to equity risk management, using principal portfolios for hedging and stress testing. Avellaneda and Lee (2010) [TODO:cite Avellaneda and Lee, 2010] applied PCA to equity returns for statistical arbitrage, exploiting mean-reversion of residuals. More recently, functional PCA methods have been applied to option-implied densities [TODO:cite Hays, Shen, and Huang, 2012] and volatility surfaces [TODO:cite Cont and da Fonseca, 2002], treating these as curves or surfaces rather than discrete vectors.

Our work differs in *what* is being decomposed: not asset returns, not yield curves, but the Fourier-cosine coefficients that encode an entire portfolio loss distribution. This is a functional PCA in spirit — each day’s observation is a 128-dimensional vector representing a density function — but executed as finite-dimensional SVD on the coefficient matrix. The key empirical finding, that 3 modes suffice for near-perfect reconstruction, parallels the yield curve result but is quantitatively stronger (97.8% vs. \$85–90% for Mode 1) and conceptually distinct (distributional dynamics rather than price dynamics).

1.4 Contribution

We show empirically that $r = 3$ suffices to capture 99.998% of all distributional variation across 250 simulated trading days:

1. **Compression:** $128 \rightarrow 3$ numbers (43×) with mean VaR error 0.005% (Section 3).
2. **Interpretation:** Mode 1 = risk level ($r = 0.998$ with VaR), Mode 2 = skew/asymmetry, Mode 3 = tail weight (Section 3.3).
3. **Dashboard:** each day is a point $(z_1, z_2, z_3) \in \mathbb{R}^3$. Anomalous days are detected by Mahalanobis distance from the cloud center (Section 3.4).
4. **Robustness:** the 3-mode structure persists out-of-sample and across portfolio sizes from 3 to 20 assets; spectral PCA outperforms raw-return PCA by a wide margin (Sections 3.6–3.7).

5. **Unification:** the learned basis is the natural state space for Bayesian filtering (#7), temporal compression (#2), Schrödinger bridges (#11), and instanton paths (#13) in the research program of Nagy (2026a, Section 7).

This is the sixth paper in a series on spectral risk: the distribution theory (Nagy, 2026a), the algorithm (Nagy, 2026b), risk measures (Nagy, 2026c), Black–Scholes verification (Nagy, 2026d), and arbitrage detection (Nagy, 2026e).

2. Method

2.1 Data Generation

We simulate $T = 250$ trading days of an $n = 5$ -asset portfolio using the Dynamic URRT framework (Nagy, 2026a, Section 7.3) with random seed 42 for full reproducibility. The simulation SDEs are:

$$d\sigma_i = 0.01 \sigma_i dW_i^{(\sigma)}, \quad d\rho_{ij} = -0.1(\rho_{ij} - \bar{\rho}_{ij}) dt + 0.005 dW_{ij}^{(\rho)}$$

where $\sigma_i(0) \in [0.15, 0.35]$ (drawn uniformly), $\bar{\rho}_{ij} = 0.4$ for all pairs, and $\rho_{ij}(0) = \bar{\rho}_{ij}$. In words:

- **Volatilities:** geometric Brownian motion with 1% daily noise
- **Correlations:** Ornstein–Uhlenbeck mean reversion ($\kappa = 0.1$) with 0.5% daily perturbation, projected to ensure positive definiteness at each step
- **Weights:** constant (equal-weight portfolio, $w_i = 1/5$)

For each day t , the Eigen-COS algorithm (Nagy, 2026b) computes $A(t) \in \mathbb{R}^{128}$ using $N = 128$ Fourier-cosine terms and integration domain $[a, b]$ determined by the cumulant-based rule of Fang and Oosterlee (2008).

2.2 SVD Decomposition

Center the coefficient matrix: $\tilde{A}(t) = A(t) - \bar{A}$. Compute the Singular Value Decomposition:

$$\tilde{A} = U\Sigma V^\top$$

where U is $T \times T$, Σ is diagonal with singular values $\sigma_1 \geq \sigma_2 \geq \dots$, and V is 128×128 . The columns of V are the principal modes v_j , and the coordinates are $z_j(t) = \tilde{A}(t)^\top v_j$.

2.3 Projection and Reconstruction

Projection: given a new coefficient vector A^* , compute

$$z_j = (A^* - \bar{A})^\top v_j, \quad j = 1, \dots, r.$$

Reconstruction: from r coordinates, recover

$$\hat{A} = \bar{A} + \sum_{j=1}^r z_j \cdot v_j.$$

The reconstruction error is $\|A^* - \hat{A}\|_2$.

3. Results

3.1 Variance Explained

Table 1 shows the cumulative variance explained by the first r modes, and Figure 1 displays the corresponding scree plot.

Modes (r)	Variance explained	Incremental
1	97.81%	97.81%
2	99.98%	2.17%
3	100.00%	0.02%
5	100.00%	$< 10^{-4}\%$

Figure 1: Scree plot. The singular values $\sigma_1, \dots, \sigma_{10}$ (log scale) show a sharp elbow after Mode 1 and negligible energy beyond Mode 3. The cumulative variance line reaches 99.998% at $r = 3$.

A single mode captures 97.8% of all distributional variation. Three modes capture 99.998% — for all practical purposes, 100%. This is a stronger result than typical PCA in finance: equity return factors typically require 5–10 components for 60–80% variance explained [TODO:cite Connor and Korajczyk, 1986], and yield curve PCA captures \$ 99% with 3 factors (Litterman and Scheinkman, 1991). The spectral coefficients are more compressible than raw asset returns because they are already a compressed representation of the distribution, as we confirm quantitatively in Section 3.7.

3.2 Reconstruction Quality

Table 2 shows the VaR reconstruction error for $r = 3$ modes. Figure 2 overlays the original VaR(5%) time series against the 3-mode reconstructed VaR.

Statistic	Coefficient L^2 error	VaR(5%) error
Mean	0.000385	0.005%
Median	0.000314	0.004%
Max	0.001858	0.042%

Figure 2: VaR reconstruction. The original VaR(5%) time series (solid) and the 3-mode reconstructed VaR (dashed) over 250 days. The two curves are visually indistinguishable; the reconstruction error (bottom panel) remains below 5 basis points throughout.

The maximum VaR error from using 3 numbers instead of 128 is 0.042% — approximately 4 basis points on a VaR of \$ \$0.80. This is two orders of magnitude below any practical materiality threshold.

3.3 Mode Interpretation

Mode 1 (z_1 , 97.8% of variance): the dominant mode correlates at $r = 0.998$ with VaR(5%) and at $r = -0.992$ with skewness. It captures the overall risk level — when z_1 increases, the distribution shifts right (higher portfolio value, lower VaR loss).

Mode 2 (z_2 , 2.2% of variance): a residual mode capturing skew-like adjustments orthogonal to the level shift. Its correlation with VaR(5%), Expected Shortfall ES(5%), and distributional skewness is low ($|r| < 0.05$ for all three), suggesting it captures a structural distributional change not easily summarized by a single risk number.

Mode 3 (z_3 , 0.02% of variance): an extremely small mode capturing fine tail adjustments. Negligible in practice.

Remark. The yield curve analogy is instructive. Three PCA modes of the Treasury yield curve are universally interpreted as “level, slope, curvature” (Litterman and Scheinkman, 1991). Our three spectral modes play an analogous role: level, asymmetry, and tail structure. The parallel is not a coincidence — both arise because smooth functions are compressible, and both the yield curve and the Fourier coefficients represent smooth objects.

Figure 3: Mode vectors. Bar charts of $v_1, v_2, v_3 \in \mathbb{R}^{128}$ showing the loading of each principal mode on the 128 Fourier coefficients. Mode 1 loads primarily on the low-frequency coefficients (A_0 through A_5), consistent with its interpretation as a level shift. Mode 2 shows an odd-symmetric pattern characteristic of skew adjustments. Mode 3 loads on mid-to-high frequencies, corresponding to tail refinements.

3.4 Anomaly Detection

The Mahalanobis distance in the 3D z -space provides a regime change detector. For the 250-day simulation, a 2.5σ threshold flags 2 anomalous days (days 227 and 230), both corresponding to a period of unusually high z_1 (elevated volatility). The detection is instantaneous — a single Euclidean distance computation in \mathbb{R}^3 .

Figure 4: 3D risk trajectory. A scatter plot of $(z_1(t), z_2(t), z_3(t))$ for $t = 1, \dots, 250$, with anomalous days (Mahalanobis distance $> 2.5\sigma$) highlighted in red. The main cluster represents normal market dynamics; the two outlier days are clearly separated, demonstrating the dashboard’s ability to flag distributional regime shifts at a glance.

Figure 5: Time series of $z_1(t), z_2(t), z_3(t)$. Three panels showing the daily evolution of each PCA coordinate over the 250-day window. z_1 (top) dominates the signal and tracks overall risk level; z_2 (middle) captures residual skew fluctuations; z_3 (bottom) is near-zero throughout, confirming that two modes account for virtually all meaningful variation.

3.5 Compression Summary

Quantity	Raw	Compressed	Ratio
Per day	128 coefficients	3 numbers	43×
250 days	32,000 numbers	1,262 numbers	25×
Information loss	—	< 0.05% VaR error	—

3.6 Out-of-Sample Validation

To test whether the learned basis generalizes beyond its training window, we split the 250-day simulation into a training set (days 1–200) and a holdout set (days 201–250). The SVD is computed on the training period only, yielding principal modes $v_1^{\text{train}}, v_2^{\text{train}}, v_3^{\text{train}}$. We then project the holdout days onto this basis and measure reconstruction quality.

Metric	In-sample (days 1–200)	Out-of-sample (days 201–250)
Variance explained (3 modes)	99.99%	99.94%
Mean VaR(5%) error	0.004%	0.009%
Max VaR(5%) error	0.035%	0.068%
Mode 1 variance share	97.9%	96.8%

The out-of-sample degradation is modest: VaR error roughly doubles but remains below 0.1%, well within practical materiality thresholds. The slight decrease in Mode 1’s variance share (97.9% → 96.8%) reflects the heightened volatility episode near days 227–230, which falls in the holdout window. The basis learned from the calm training period still captures the holdout dynamics, including the anomalous days.

3.7 Robustness Analysis

We test sensitivity to three key design choices: portfolio size, correlation regime, and crisis dynamics.

Portfolio size. We repeat the full analysis with $n \in \{3, 5, 10, 20\}$ assets, holding all other simulation parameters constant.

Assets (n)	Mode 1 variance	3-mode cumulative	Max VaR error (3 modes)
3	98.4%	100.00%	0.028%
5	97.8%	99.998%	0.042%
10	96.1%	99.95%	0.11%
20	93.7%	99.82%	0.24%

As the portfolio grows, the coefficient dynamics become higher-dimensional: with 20 assets, Mode 1 drops to 93.7% and three modes capture 99.82% rather than 100%. This is expected — more assets introduce more independent sources of distributional variation. Nevertheless, the 3-mode compression remains effective (VaR error below 0.25%) even at $n = 20$. For larger portfolios, a 4th or 5th mode may be warranted.

Correlation regime. We vary the mean correlation $\bar{\rho}$ across $\{0.0, 0.2, 0.4, 0.6, 0.8\}$. Higher correlation concentrates variance into fewer modes: at $\bar{\rho} = 0.8$, Mode 1 captures 99.3% of variance, and the coefficient dynamics are nearly one-dimensional. At $\bar{\rho} = 0.0$ (independent assets), Mode 1 drops to 91.2% and five modes are needed for 99.9%. The intuition is straightforward: correlated assets move together, so their spectral coefficients evolve along fewer directions.

Crisis regime. Using the `simulate_2008_crisis` function in the codebase (which introduces a sudden correlation spike to $\rho = 0.9$ and a volatility doubling at day 125), we test whether the pre-crisis basis detects the regime change. The result is unambiguous: the Mahalanobis distance in

(z_1, z_2, z_3) -space spikes to $> 5\sigma$ on the crisis onset day, and 3-mode reconstruction error jumps from 0.01% to 1.8% VaR. This confirms two points: (i) the 3D basis is an effective anomaly detector, because the crisis manifests as a large outlier in z -space, and (ii) the basis requires re-estimation after a structural break, as anticipated in Section 6.

Comparison: spectral PCA vs. raw-return PCA. A natural baseline is to apply PCA directly to daily portfolio returns (a $T \times n$ matrix) rather than to spectral coefficients (a $T \times 128$ matrix). For $n = 5$ assets and $T = 250$ days, PCA on raw returns yields 3-mode cumulative variance of 89.4%, compared to 99.998% for spectral PCA. This gap arises because spectral coefficients encode the *entire loss distribution*, not just the first moment (return). Distributional variation is smoother and more compressible than return variation, because the Fourier basis already imposes regularity.

4. The Unification Property

The learned basis is not just a compression tool. It is the **natural coordinate system** for the entire spectral risk research program:

4.1 Dynamic URRT (Nagy, 2026a, Direction 2)

The temporal compression of spectral coefficients is a PCA problem. With the learned basis, the Dynamic URRT reduces to tracking $z_1(t), z_2(t), z_3(t)$ instead of 128 coefficients. The SVD rank of the coefficient time series equals the number of learned basis modes: both are 1–3.

4.2 Bayesian Live Risk (Nagy, 2026a, Direction 7)

The Bayesian posterior on the 128 coefficients becomes a posterior on 3 numbers:

$$P(z_1, z_2, z_3 \mid \text{data}).$$

A 3D Kalman filter replaces a 128D filter. The computational cost drops by a factor of $(128/3)^3 \approx 77,000$.

4.3 Schrödinger Bridge (Direction 11)

The minimum-entropy path between two distributions is a path in \mathbb{R}^3 :

$$z(s), \quad s \in [t, t+1]$$

connecting today's (z_1, z_2, z_3) to tomorrow's. The 128-dimensional optimal transport problem becomes 3-dimensional.

4.4 Instanton Paths (Direction 13)

The most likely crash scenario is a path in z -space from the current state to a high-VaR state. The Euler–Lagrange equations are 3D ODEs, not 128D.

5. Relationship to Yield Curve PCA

The closest analogy in finance is the PCA decomposition of interest rate yield curves (Litterman and Scheinkman, 1991; Dai and Singleton, 2000). Both share the same structure:

Property	Yield curve PCA	Spectral coefficient PCA
Object	Bond yields $y(T)$	Spectral coefficients A_k
Dimension	\$ \$10 maturities	128 modes
Modes needed	3 (level, slope, curve)	3 (level, skew, tail)
Variance by Mode 1	\$ \$85–90%	97.8%
Variance by 3 modes	\$ \$99%	\$ \$100%
Physical interpretation	Interest rate dynamics	Risk distribution dynamics

The spectral coefficients are **more compressible** than yield curves because they are already a Fourier basis — a basis optimized for representing smooth functions. Smooth functions in a smooth basis are maximally compressible.

6. The 130 vs 3 Distinction

A natural question arises: if 3 numbers suffice, why do we need 130?

The answer is that the two representations serve fundamentally different purposes:

Property	130 parameters (URRT)	3 parameters (PCA)
What it encodes	Any distribution, from scratch	Change relative to a known baseline
Prerequisites	None (portfolio data only)	250 days of training + stored basis
Guarantee	Mathematical (Theorem 7, Nagy 2026a)	Empirical (conditional on training regime)
Failure mode	Never (unconditional)	Regime change outside training range
Analogy	GPS coordinates (any point on Earth)	“3 blocks left” (requires knowing where you are)

The URRT’s 130-parameter bound is **unconditional**: it holds for any portfolio, any correlation structure, any volatility profile, without prior data. The PCA compression to 3 numbers is **conditional** on the training period being representative of future market dynamics. In a regime change (2008-type crisis), the learned basis becomes stale and the full 130 parameters are needed to represent the new distributional state.

The correct interpretation is hierarchical:

1. The **URRT** establishes that 130 numbers are sufficient for any single snapshot (static).

2. The **PCA** reveals that within a regime, these 130 numbers typically move along 3 directions (dynamic).
3. When the regime changes, the 3-dimensional trajectory exits the learned subspace, signaling that the basis needs re-estimation — itself an anomaly detection mechanism.

This hierarchy is the bridge to Bayesian Live Risk (Direction #7): the 3 PCA coordinates are the state space for the Kalman filter, and the width of the Kalman posterior detects when the basis is losing grip (i.e., when the 3-dimensional model is no longer adequate and the full 130-dimensional state must be consulted).

7. Lean 4 Verification

We formalize the mathematical foundations of PCA compression in Lean 4. Transparency requires distinguishing between results that are *fully proved* from Mathlib primitives and results that are *stated and type-checked* with their core mathematical content assumed as hypotheses. Table 4 reports both categories.

Fully proved results (non-trivial proofs from Lean/Mathlib primitives):

Result	Lean file	Status
Variance ratio $\in [0, 1]$	LearnedBasis.lean	Proved
3-mode sufficiency criterion	LearnedBasis.lean	Proved
Additional mode increases variance	LearnedBasis.lean	Proved
Parseval’s identity (\sum -form for finite orthonormal systems)	Parseval.lean	Proved

These four results are verified end-to-end: they derive their conclusions from definitions and Mathlib lemmas with no assumptions beyond the stated hypotheses. The sufficiency criterion, for instance, shows that if the cumulative variance ratio exceeds a threshold τ , the reconstruction error is bounded by $(1 - \tau)\|\tilde{A}\|^2$.

Axiomatized results (statement type-checked, core content assumed):

Result	Lean file	Status
Bessel’s inequality ($\sum z_j^2 \leq \ A\ ^2$)	LearnedBasis.lean	Axiomatized
Eckart–Young optimality	EckartYoung.lean	Axiomatized
VaR reconstruction bound	LearnedBasis.lean	Axiomatized
Mahalanobis threshold bound	LearnedBasis.lean	Axiomatized

These four results are correctly *stated* and type-check in Lean 4, but their proofs assume the core mathematical content as a hypothesis and return it unchanged. They serve as formal specification (the statement is machine-verified to be well-typed), not as formal proof. Full formalization of

Bessel’s inequality and Eckart–Young from Mathlib’s inner product space library is in progress; we state them as axioms here to make the dependency structure explicit.

All eight files compile with zero sorry. The distinction between proved and axiomatized results is important: the proved results suffice to guarantee the variance ratio bounds and sufficiency criterion used in Sections 3.1 and 3.5. The axiomatized results (Bessel, Eckart–Young) provide the theoretical optimality guarantees; their textbook proofs are standard [TODO:cite Golub and Van Loan, 2013] but their Lean formalization requires matrix analysis infrastructure not yet available in Mathlib.

8. Conclusion

The risk profile of a portfolio, encoded as 128 spectral coefficients, is effectively three-dimensional over time. This is an empirical fact, not an assumption: PCA on the coefficient time series reveals that a single mode captures 97.8% of all variation, and three modes capture 99.998% (Table 1, Section 3.1).

The practical implication is a 3-number risk dashboard:

- z_1 = how much the overall risk level changed
- z_2 = how much the asymmetry changed
- z_3 = how much the tail weight changed

Any day’s full distributional change — including all moments, all quantiles, VaR, ES, and any spectral risk measure — is captured by these three coordinates, with reconstruction error below 0.05%.

The learned basis is the unifying computation for the spectral risk program: Dynamic URRT, Bayesian filtering, optimal transport, and instanton analysis all reduce to 3-dimensional problems once the basis is extracted.

Limitations. The empirical results are based on simulated data with a specific parametric structure (Section 2.1). While the out-of-sample and robustness analyses in Sections 3.6–3.7 demonstrate stability across holdout periods and portfolio sizes, validation on real market data remains essential future work. The 3-mode sufficiency is conditional on the training regime being representative; in a structural break outside the training distribution, the full 130-parameter representation (Section 6) should be used until the basis is re-estimated. Finally, four of the eight Lean-formalized results are axiomatized rather than fully proved (Section 7), and completing these formalizations is ongoing work.

During the preparation of this work the author used large language models in order to assist with manuscript drafting, literature search, and coding assistance. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

- Dai, Q. and Singleton, K.J (2000). Specification analysis of affine term structure models. *Dai, Q. and Singleton, K.J.*, 55(5). DOI: 10.3386/w6128
- Fang, Fang and Oosterlee, Cornelis W. (2008). A Novel Pricing Method for European Options Based on Fourier-Cosine Series Expansions. *SIAM Journal on Scientific Computing*, 31(2), 826-848. DOI: 10.1137/080718061
- Litterman, R. and Scheinkman, J (1991). Common factors affecting bond returns. *Litterman, R. and Scheinkman, J.*, 1(1). DOI: 10.3905/jfi.1991.692347
- Nagy, T. (2026). The Fenton Distribution Solved (with Latent) - An Elementary CDF for Sums of Correlated Lognormals. *Zenodo*. DOI: 10.5281/zenodo.19144775
- Fang, F. and C. W. Oosterlee (2009). COS method. *SIAM J. Sci. Comput.*, 31(2).
- Nagy, T. (2026). Noise-Free Risk: Deterministic VaR, ES, and Spectral Risk Measures for Lognormal Portfolios. *Working paper*.
- Nagy, T. (2026). From Itô to Black–Scholes: A Machine-Verified Derivation in Lean 4. *Zenodo*. DOI: 10.5281/zenodo.18910551
- Nagy, T. (2026). The Anomaly Functional: Real-Time Arbitrage Detection via Spectral Risk Coefficients. *Working paper*.