Supplementary Material:
Information-Theoretic Disentanglement and Evaluation of Deep Representations in Quantitative Finance

Zhaozhao Ma[⋆,†]
[⋆]*Zhejiang University*
[†]*Georgia Institute of Technology*
zhaozhaoma@{zju.edu.cn, gatech.edu}

Table of Contents

A Supplementary Explanations on Framework Implementation

A.1 Overall Framework

As shown in Fig. 1, the framework processes heterogeneous inputs(e.g., $\mathbf{X}_1$: Price, $\mathbf{X}_2$: News) through three phases:

(a) Geometric anchoring: Establishes ground truth topological baseline via fixed statistical anchors $\mathbf{U}$ alongside frozen deep embeddings $\mathbf{Z}$, which ensures that subsequent evaluations are grounded in the intrinsic geometry of the raw data.

(b) Separate variational bottleneck distillation: Employs learnable bottleneck probes $\mathbf{V}_1$ and $\mathbf{V}_2$ to actively distill task relevant signals from high-dimensional noise by maximizing geometric dependence L_Y respectively, while suppressing residuals. This yields a geometric sufficient statistic for each source.

(c) PID disentanglement: Project the distilled representations $\mathbf{V}_1$ and $\mathbf{V}_2$ into a joint geometric space. Using subspace alignment, we decompose their predictive power into Shared Information (redundancy) and Unique Alpha, providing a rigorous, quantitative basis for model selection or ensemble weighting.

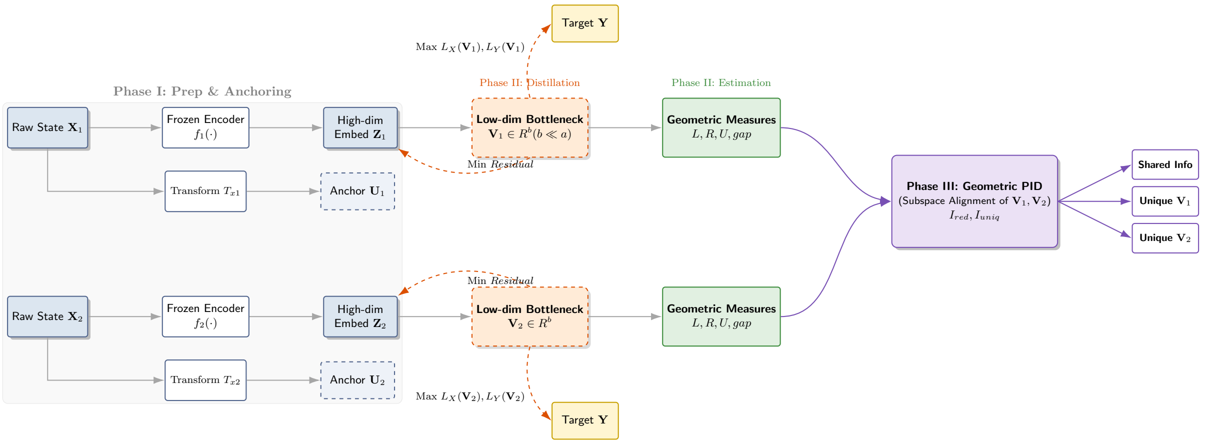

Figure 1: Overall workflow of the proposed method.

A.2 Motivation of the Geometric Anchor U

The fixed representation $\mathbf{U}$, introduced in Eq. 1, serves as a topological definition of the input space for data valuation. Its role is twofold:

- **Provide Input Kernel:** Since raw inputs X (unstructured text or sparse arrays) lack a canonical Hilbert space structure, $\mathbf{U}$ provides a standardized low-dimensional manifold to construct the input kernel matrix $\tilde{\mathbf{K}}_X$. This allows the input dependence measure $L_X = \hat{I}_{\text{dep}}(\tilde{\mathbf{K}}_X, \tilde{\mathbf{K}}_V)$ to be rigorously computed during bottleneck learning.
- **Null-Hypothesis for Valuation:** $\mathbf{U}$ also enables a principled post-hoc valuation, by comparing the predictive dependence of the learned bottleneck $L_Y(V)$ against that of the shallow anchor $L_Y(\mathbf{U})$, we can verify if the deep model captures genuine non-linear alpha ($L_Y(V) \gg L_Y(U)$) or merely memorizes trivial statistical artifacts.

A.3 Motivation of Optimization Objective

The learning objective in Phase II (Eq. 12) is designed to act as Information sufficient statistics learner. We simultaneously maximize principal terms while minimizing residuals:

- Maximize $L_Y(\mathbf{V})$ to Ensure predictive sufficiency, forcing the bottleneck to capture as much target information as possible.
- Maximize $L_X(\mathbf{V})$ to ensure structural fidelity. By maximizing dependence with the anchor $\mathbf{U}$, we constrain the bottleneck to retain the intrinsic topological structure of the data modality, preventing the representation from drifting into spurious feature spaces.
- Minimize R_Y, R_X to ensure completeness. Minimizing residuals guarantees that the information captured in L is a tight lower bound of the true information content.

B Evaluation Metrics

For each evaluation, in addition to reporting lower and upper bounds, residuals, and gap metrics, uncertainty is quantified via bootstrap and permutation methods, yielding point estimates and confidence intervals:

Uncertainty via bootstrap. Report point estimates together with $(1-\alpha)$ confidence intervals for L_s, U_s, gap_s using nonparametric bootstrap.

Permutation calibration. Define $M_s = \frac{1}{2}\big(L_s + U_s\big)$. Under a permutation null, estimate (μ_0, σ_0) of M_s and report z-scores $z_s = (M_s - \mu_0)/\sigma_0$, lower-tailed p-values, and a calibrated index:

$$\text{CI}_s = \text{clip}\left(\frac{\mu_0 - M_s}{\mu_0 - L_s},\ 0,\ 1\right), \tag{1}$$

which maps null-like performance to 0 and the lower bound to 1.

C Planned Experiments and Expected Outcomes

To validate both the theoretical properties and practical utility of the proposed framework, this proposal adopts a two-stage evaluation protocol.

C.1 Experiment 1: Synthetic Data Validation

Setup. We will simulate non-stationary financial time series via a Regime-Switching Heston Model Heston [1993], Hamilton [1989] alternating between Regime A (low volatility, mean-reverting) and Regime B (high volatility, momentum-driven), with signal-to-noise ratios (SNR) varying from 0.1 to 2.0. A standard TCN Bai et al. [2018] serves as encoder $f(\cdot)$, while a simple AR model defines the geometric anchor $\mathbf{U}$ as a shallow, interpretable baseline.

- **$\mathbf{H}_1$:** $L_Y(V)$ exhibits strict positive monotonicity with SNR, while the gap $= U_{\text{bound}} - L$ expands as SNR decreases, quantifying epistemic uncertainty.

C.2 Experiment 2: Real-Market Multimodal Disentanglement

Setup. We will construct a multimodal dataset for S&P 500 constituents, temporally aligning high-frequency LOB snapshots Huang and Polak [2011] with timestamped financial news headlines (like Reuters or Dow Jones). Price modality employs whitened PCA as anchor $\mathbf{U}_{\text{price}}$ and DeepLOB Zhang et al. [2019] (CNN–LSTM) as encoder f_{price}; news modality uses TF-IDF as anchor $\mathbf{U}_{\text{news}}$ and FinBERT Araci [2019] as encoder f_{news}. The prediction target is 1-to-10-minute forward active returns.

- **$\mathbf{H}_2$ (Event-Driven Flow):** Unique news information $I_{\text{uniq}}(\text{news})$ peaks during high-entropy corporate events (earnings, FOMC), while redundant information I_{red} dominates quiet periods.
- **$\mathbf{H}_3$ (Market Efficiency):** Liquid large-caps exhibit low $I_{\text{uniq}}(\text{news})/I_{\text{total}}$ ratios (high efficiency), while small-caps show significantly higher ratios, revealing cross-sectional informational asymmetry.
- **$\mathbf{H}_4$ (PID-Weighted Ensemble):** A PID-weighted ensemble, allocating weights by I_{uniq}, achieves superior Information Ratios versus equal-weighted baselines, translating information decomposition into actionable alpha.

D Statistical Properties of $\widehat{I}_{\text{dep}}$

To validate $\widehat{I}_{\text{dep}}$ as a reliable statistical measure, we need additionally apply standard statistical theory to derive its consistency, bias, and asymptotic normality.

Population Quantities. Let $k_c(x, x')$ be the centered kernel, define the population-level quantities (targets of the V-statistics):

$$\mu_{KL} := \mathbb{E}[k_c(X, X')\ell_c(Z, Z')], \quad \mu_K := \mathbb{E}[k_c(X, X')^2], \quad \mu_L := \mathbb{E}[\ell_c(Z, Z')^2]$$

The true population target is $I^*_{\text{dep}} = -\log\left(\frac{\mu_{KL}^2}{\mu_K \mu_L}\right)$.

Lemma 1 (V-Statistic Consistency [Serfling, 2009]). *Assuming bounded and second-order integrable kernels, the normalized Frobenius inner products are V-statistics that converge almost surely to their population counterparts:*

$$\frac{1}{n^2}\langle \tilde{\mathbf{K}}, \tilde{\mathbf{L}} \rangle_F = \frac{1}{n^2}\sum_{i,j} k_c(X_i, X_j)\ell_c(Z_i, Z_j) \xrightarrow{a.s.} \mu_{KL}$$

$$\frac{1}{n^2}\langle \tilde{\mathbf{K}}, \tilde{\mathbf{K}} \rangle_F \xrightarrow{a.s.} \mu_K, \quad \frac{1}{n^2}\langle \tilde{\mathbf{L}}, \tilde{\mathbf{L}} \rangle_F \xrightarrow{a.s.} \mu_L$$

Proof. This follows directly from the strong law of large numbers for V-statistics [Serfling, 2009]. □

Lemma 2 (Consistency). *Under the conditions of Lemma 1 and assuming non-degenerate kernels ($\mu_K, \mu_L > 0$), the estimator is strongly consistent:*

$$\widehat{I}_{dep}(\mathbf{K}, \mathbf{L}) \xrightarrow{a.s.} I^*_{dep}$$

Proof. This follows directly Lemma 1 and the continuous mapping theorem. □

Lemma 3 (Finite Sample Bias). *Assume bounded and fourth-order integrable kernels. Let $I_n = \widehat{I}_{dep}(\mathbf{K}, \mathbf{L})$. The bias of the estimator is of order $O(n^{-1})$:*

$$\left|\mathbb{E}[I_n] - I^*_{dep}\right| \leq \frac{C}{n} + o\left(\frac{1}{n}\right)$$

Proof. Let $A_n = \frac{1}{n^2}\langle \tilde{K}, \tilde{L}\rangle_F$, $B_n = \frac{1}{n^2}\langle \tilde{K}, \tilde{K}\rangle_F$, $C_n = \frac{1}{n^2}\langle \tilde{L}, \tilde{L}\rangle_F$, $V_n = (A_n, B_n, C_n)^\top$, with population limits $\mu = (\mu_{KL}, \mu_K, \mu_L)^\top$. Each of A_n, B_n, C_n is a second order V-statistic, standard results give us:

$$\mathbb{E}[V_n] - \mu = O(n^{-1}).$$

Define $g(a, b, c) = -\log\big(a^2/(bc)\big)$ and $I_n = g(V_n)$, $I^* = g(\mu)$. A first order mean value expansion yields:

$$I_n - I^* = \nabla g(\mu)^\top (V_n - \mu) + R_n, \qquad \nabla g(\mu) = \big[-2/\mu_{KL},\, 1/\mu_K,\, 1/\mu_L\big]^\top,$$

where $\mathbb{E}|R_n| = o(n^{-1})$ under the stated moment and non–degeneracy assumptions. Taking expectations,

$$\big|\mathbb{E}[I_n] - I^*\big| \leq \|\nabla g(\mu)\|\, \big\|\mathbb{E}[V_n] - \mu\big\| + o(n^{-1}) = O(n^{-1}).$$

□

Lemma 4 (Asymptotic Normality). *Under the conditions of Lemma 3, the estimator is asymptotically normal:*

$$\sqrt{n}\left(\widehat{I}_{dep}(\mathbf{K}, \mathbf{L}) - I^*_{dep}\right) \xrightarrow{d} \mathcal{N}\left(0, \nabla g(\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma} \nabla g(\boldsymbol{\mu})\right)$$

where $\boldsymbol{\mu} = [\mu_{KL}, \mu_K, \mu_L]^\intercal$, $\nabla g(\boldsymbol{\mu}) = [-2/\mu_{KL}, 1/\mu_K, 1/\mu_L]^\intercal$, and $\boldsymbol{\Sigma}$ is the covariance matrix of the Hoeffding projections.

Proof. Using the notation of Lemma 3, the vector of second order V-statistics satisfies joint CLT $\sqrt{n}\,(V_n - \mu) \Rightarrow \mathcal{N}(0, \Sigma)$. Since g is C^1 at μ, the multivariate Delta method yields $\sqrt{n}\,(g(V_n) - g(\mu)) \Rightarrow \mathcal{N}(0, \nabla g(\mu)^\top \Sigma \nabla g(\mu))$. □

E Formal Justification of Geometric Residuals

To validate geometrically defined Eq.8 is a valid and convergent statistical quantity, requires showing our sample based RFF projector Π converges to the correct population level operator.

Lemma 5 (RFF Projector Consistency). *Let P_V be the true population level RKHS projection operator onto the subspace spanned by V. Let $\Pi = \Pi_{n,m}$ be the sample based projector from Eq.6. If m is chosen appropriately (e.g., $m \gtrsim \sqrt{n}\log n$ [Rudi and Rosasco, 2017, Sutherland and Schneider, 2015]), then the sample projector satisfy:*

$$\|\Pi_{n,m} - P_V\|_{op} \xrightarrow{p} 0 \quad \text{as } n, m \to \infty.$$

Let $\varepsilon_n = \|(\Pi - P_V)\mathbf{K}_Z\|_F / \|\mathbf{K}_Z\|_F$. The conditions above imply $\varepsilon_n = o_p(1)$.

Proof. This result is a direct consequence of RFF kernel approximation [Rahimi and Recht, 2007] and kernel regression [Rudi and Rosasco, 2017]. □

Theorem 1 (Residual Estimator Convergence). *Under the conditions of Lemma 5, the residual estimator $\widehat{I}_{dep}(\mathbf{K}_X, \tilde{\mathbf{G}}_{res})$ converges in probability to its population level target, the finite sample error is bounded:*

$$\left|\widehat{I}_{dep}(\mathbf{K}_X, \tilde{\mathbf{G}}_{res}) - I^*_{\text{dep, pop-residual}}\right| = O_p(\varepsilon_n + n^{-1/2})$$

Proof. The total error is a sum of the statistical (V-stat) error ($O_p(n^{-1/2})$) and the projector approximation error (ε_n) from Lemma 5. □

Proof. Recall, we have $\|\Pi_\lambda - P_V\|_{\mathrm{op}} \xrightarrow{p} 0$, hence $M_\lambda = I - \Pi_\lambda \to P_V^\perp$ and

$$\|\tilde{G}_{\mathrm{res}} - \widetilde{P_V^\perp K_Z P_V^\perp}\|_F \;\le\; C\,\|(\Pi_\lambda - P_V)K_Z\|_F \;=\; O_p(\varepsilon_n).$$

Let $F(B) := -\log\big(\langle \tilde{K}_X, \tilde{B}\rangle_F^2 / (\langle \tilde{K}_X, \tilde{K}_X\rangle_F \langle \tilde{B}, \tilde{B}\rangle_F)\big)$. On $\{\mu_X > 0, \mu_{\mathrm{res}} > 0, \mu_{X,\mathrm{res}} > 0\}$ the map F is C^1 in a neighborhood of $B_\star = \widetilde{P_V^\perp K_Z P_V^\perp}$, so by a mean–value bound

$$\big|F(\tilde{G}_{\mathrm{res}}) - F(B_\star)\big| = O_p(\varepsilon_n).$$

With $B_\star$ fixed, $F(B_\star)$ is a smooth function of order-2 V-statistics, hence by the joint CLT and the Delta method:

$$F_n(B_\star) - F(B_\star) = O_p(n^{-1/2}).$$

Combining the two displays gains $\big|\widehat{I}_{dep}(K_X, \tilde{G}_{\mathrm{res}}) - I^*_{dep,\mathrm{pop\text{-}res}}\big| = O_p(\varepsilon_n + n^{-1/2})$. ☐

F Upper Bound Completeness

Proof. Let:

$$c := \mathrm{CKA}(A,B) = \frac{|\langle \widetilde{A}, \widetilde{B}\rangle_F|}{\|\widetilde{A}\|_F \|\widetilde{B}\|_F}, \quad c_1 := \mathrm{CKA}(A, B_\parallel), \quad c_2 := \mathrm{CKA}(A, B_\perp),$$

and set $\kappa = \|\widetilde{A}\|_F$, $a = \|\widetilde{B}_\parallel\|_F$, $b = \|\widetilde{B}_\perp\|_F$, $x = |\langle \widetilde{A}, \widetilde{B}_\parallel\rangle_F| = \kappa a\, c_1$, $y = |\langle \widetilde{A}, \widetilde{B}_\perp\rangle_F| = \kappa b\, c_2$. by triangle inequality:

$$c = \frac{x+y}{\kappa\|\widetilde{B}_\parallel + \widetilde{B}_\perp\|_F} \;\ge\; \frac{x+y}{\kappa(a+b)} = \frac{c_1 a + c_2 b}{a+b} \;\ge\; \min\{c_1, c_2\}.$$

Applying the monotone map $-2\log(\cdot)$ yield:

$$\widehat{I}_{dep}(A,B) \;\le\; \max\{-2\log c_1,\, -2\log c_2\} = \max\{L_X, R_X\}.$$

Since $\max\{u,v\} \le \log(e^u + e^v)$ for all $u, v \in \mathbb{R}$, we can obtain:

$$\widehat{I}_{dep}(A,B) \;\le\; \log\big(e^{L_X} + e^{R_X}\big).$$

☐

References

Steven L Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6(2):327–343, 1993.

James D Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Ruihong Huang and Tomas Polak. Lobster: Limit order book reconstruction system. *Available at SSRN 1977207*, 2011.

Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*, 67(11):3001–3012, 2019.

Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017.

Danica J Sutherland and Jeff Schneider. On the error of random fourier features. *arXiv preprint arXiv:1506.02785*, 2015.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pages 1177–1184, Vancouver, Canada, 2007.