INTERPRETABLE HETEROGENEOUS INSTRUMENTED PRINCIPAL COMPONENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Instrumented principal component analysis (IPCA) is an important category of PCA that is anchored to a set of instruments (e.g., treatments). However, standard IPCA neither supports heterogeneity nor extracts interpretable factors. To solve these problems, we propose a novel interpretable heterogeneous IPCA (HeterIPCA). It constructs the factor loadings with rich heterogeneity by a structured graph convolutional network (SGCN), achieves a nonasymptotic bound of parameter estimation, and establishes asymptotic normality of inference. Experimental results show that HeterIPCA achieves state-of-the-art performance on various evaluation metrics. This finding may shed new light on the interpretability and heterogeneity of PCA and factor analysis.

1 INTRODUCTION

Principal component analysis (PCA, Vaswani et al. 2018) has been studied for more than a century as a common methodology for dimension reduction, feature extraction, and outlier removal. The standard PCA works in an unsupervised way by exploiting the covariance structure of data. Recently, a supervised version of PCA that is anchored to a set of instruments has attracted more and more attention, named instrumented PCA (IPCA, Kelly et al. 2019; Gu et al. 2020; Leippold et al. 2022; Chen et al. 2024; Xiang et al. 2024; Duan et al. 2025; Lai & Yang 2025). In this framework, the instruments can be treatments, characteristics regarding background knowledge, or other observable factors. This scheme helps consistently estimate the latent factors and recover the loadings associated with these factors.

However, one main drawback of IPCA is the non-uniqueness of factorization, which causes multiple sets of factors and loadings. This undermines the interpretability and explainability of IPCA models, although IPCA itself aims to reveal underlying factors in machine learning tasks. To fix this problem, a standard treatment is to impose several constraints on the factors and loadings, such as orthogonal loadings, diagonal second moment of factors, and nonnegativity (or nonpositivity) (Kelly et al., 2019; Davis et al., 2025; Duan et al., 2025). But even these treatments cannot narrow down the ambiguity of factorization to a satisfactory region. Hence these resulted factors are called **weak factors** by (Davis et al., 2025). Another drawback of IPCA is the **lack of heterogeneity**, since it assumes a shared global mapping between instruments and factor loadings. In reality, the relationship between an instrument and a factor loading can be highly context-dependent, influenced by underlying issues like the industry of a firm, the demographics of a customer, or unique corporate events. IPCA fails to accommodate such nuanced and individual-level variations (Lai & Wang, 2024; Wang et al., 2025; Lai et al., 2026a;b) outside of the common and linear mapping. Consequently, it has to discard such heterogeneity into the residual term, which is intentionally used to capture the unexplainable portion of the model.

Recently, deep neural networks (DNNs) have been adopted to improve nonlinear explanatory power and individual heterogeneity, such as the Conditional Autoencoder (CA) (Gu et al., 2021), the Factor Graph Contrastive Learning (FactorGCL), and the Structured DNN (SDNN) (Farrell et al., 2025). However, CA and FactorGCL **lack theoretical guarantees** for the first stage on **parameter estimation**, needless to say the second stage on **statistical inference**. Although SDNN theoretically supports parameter estimation and statistical inference, the current theoretical results are **restricted to** the simple network class: **fully connected networks (FCN)**. Worse still, it has been empirically

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

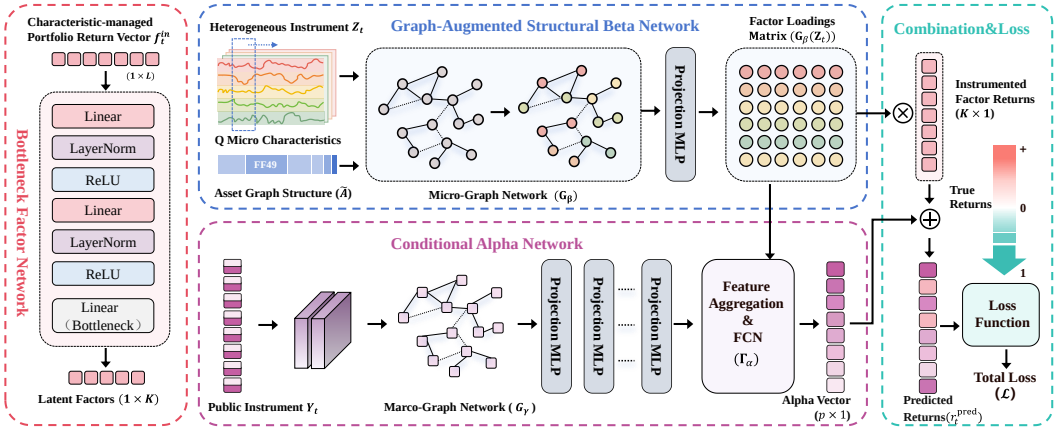


Figure 1: Architecture of the proposed HeterIPCA, which consists of four main components: **(a) Micro-Graph Network** G_β processes the heterogeneous instruments Z_t (individual-level characteristics) and the asset graph structure $\tilde{A}$ to generate a dynamic factor loading matrix. **(b) Macro-Graph Network** G_γ captures the influence of public instruments Y_t. **(c) Conditional Alpha Network** Γ_α fuses G_β and G_γ to estimate the alpha vector (abnormal returns). **(d) Bottleneck Factor Network** extracts K latent factors and combines them with the factor loadings to predict returns r_{pred}.

implemented only for a **single** $\{0, 1\}$**-discrete treatment**, which is **far from enough** in real-world scenarios with multiple continuous treatments.

The above unsolved problems motivate us to develop an interpretable Heterogeneous IPCA (**HeterIPCA**), in order to: **1)** enrich the parameters of IPCA into more flexible functions of observables; and **2)** extend the linear representation of IPCA to the more flexible non-linear representation. Intuitive as it is, there are several significant difficulties to be addressed:

1. It is a **non-standard statistical inference** for the parameter functions of DNN in the IPCA setting.

2. A new problem emerges that how to **identify factors** for IPCA via the non-linear DNN instrumenting. This is crucial for preserving the econometric interpretation of HeterIPCA.

3. How to retrieve **both global and context-dependent** relationships between instruments and factor loadings remains unsolved.

Our main contributions fall into the following aspects, which overcome the above difficulties:

1. We reformulate IPCA by a structured graph convolutional network (SGCN), which **theoretically supports statistical inference** for meaningful quantities with **asymptotic normality**.

2. We directly use the parameter functions as factor loadings for IPCA and calculate the **pseudo-dimension of SGCN**, which enable well-defined factor interpretation and **parameter estimation** with a sharp enough **nonasymptotic bound**.

3. We **embed heterogeneity** into the SGCN, which preserves **both global and context-dependent correlations** and improves explanatory power of the model.

The entire architecture of HeterIPCA and a brief explanation are provided in Figure 1.

2 Preliminaries and Related Work

2.1 Instrumented Principal Component Analysis

The IPCA model (Kelly et al., 2019) integrates the factor model with observable instruments, shown as follows:

$$r_{t+1} = Z_t\Gamma_\alpha + Z_t\Gamma_\beta f_{t+1} + \epsilon_{t+1}, \tag{1}$$

where r_{t+1} is a $p \times 1$ vector of dependent variables, Z_t is a $p \times \mathscr{L}$ matrix that stacks the $\mathscr{L}$ instruments for the p dependent variables, Γ_α is an $\mathscr{L} \times 1$ vector that describes the relationship between observable instruments and the abnormal intercept, Γ_β is an $\mathscr{L} \times K$ factor loading matrix that describes the relationship between observable instruments and the K factors f_{t+1}, and ϵ_{t+1} is a composite error. It can be directly seen that IPCA uses a linear representation of factor loadings $Z_t\Gamma_\beta$. This setting has a widely-recognized drawback that there are infinite rotation-equivalent solutions $\Gamma_\beta R^{-1}$ and Rf_{t+1} of factor loadings and factors, respectively, where R is any rotation matrix. This drawback undermines the interpretability and explainability of IPCA models. To address

this drawback, IPCA further imposes several constraints like $\Gamma_\beta^\top \Gamma_\beta = \mathbb{I}_K$ and diagonal second moment. But even this treatment cannot narrow down the ambiguity of factorization to a satisfactory region.

2.2 Conditional Autoencoder Asset Pricing Model

By recognizing the drawback of linear factor loadings, (Gu et al., 2021) further propose an extended nonlinear asset pricing model via the Autoencoder (Hinton & Salakhutdinov, 2006), named conditional Autoencoder (CA):

$$r_{t+1} = \Gamma_\beta(Z_t) r_t + \epsilon_{t+1}, \tag{2}$$

where $\Gamma_\beta(Z_t)$ is the factor loadings built by the Autoencoder with input Z_t, and the factors f_{t+1} are set to the asset returns r_t at time t. In this way, CA works as a standard autoregressive model that aims to predict asset returns at time $(t+1)$ with asset returns and instruments at time t. Compared with the original IPCA, CA improves flexibility in modeling factor exposures. However, it lacks not only theoretical guarantees for parameter estimation, but also a specialized mechanism to effectively capture high-order complex dependencies between asset returns and factors, which requires a more advanced network structure.

2.3 Factor Graph Contrastive Learning

The Factor Graph Contrastive Learning (FactorGCL, Duan et al. 2025) employs a cascading residual hypergraph architecture to complement human-designed prior factors with data-driven hidden factors. It further models the residual ϵ_{t+1} in (2) by a HyperGraph Convolutional neural Network (HyperGCN) and extend (2) to the following FactorGCL model:

$$r_{t+1} = G_\beta(Z_t, r_t) + G_\alpha(Z_t, r_t), \tag{3}$$

where $G_\beta(\cdot,\cdot)$ and $G_\alpha(\cdot,\cdot)$ are two HyperGCNs that simultaneously process both the instruments Z_t and historical asset returns r_t to predict the next asset returns r_{t+1}. However, **this setting drops both the expressiveness of IPCA and the consistency of statistical inference**. The form of factor loadings $G_\beta(Z_t)$ is lost, which invalidates the rigorous form of statistical inference and empties consistent estimations of model parameters that have clear practical meanings. In other words, FactorGCL works as a standard deep learning (DL) method that receives Z_t and r_t, and then outputs r_{t+1}, without a rigorous statistical inference of network weights, or the network weights lack clear practical meanings.

2.4 Structured Deep Neural Networks for Individual Heterogeneity

The interpretability problem of DL has been intensively addressed to improve reliability of applying such models. Recently, SDNNs have been found to be highly interpretable for modeling individual heterogeneity (Farrell et al., 2025). In the context of our task, it allows for varying model parameters $\Gamma(\cdot)$ specified at individual instruments Z, instead of fixed model parameters Γ:

$$\Gamma^*(\cdot) = \underset{\Gamma \in \mathscr{F}_{DNN}}{\operatorname{argmin}} \ \mathbb{E}[\ell(r, f, \Gamma(Z))], \tag{4}$$

$$\hat{\Gamma}(\cdot) = \underset{\Gamma \in \mathscr{F}_{DNN}}{\operatorname{argmin}} \ \frac{1}{T} \sum_{t=0}^{T-1} \ell(r_{t+1}, f_{t+1}, \Gamma(Z_t)), \tag{5}$$

where $\mathscr{F}_{DNN}$ denotes a certain type of DNNs, and ℓ denotes the loss function involving dependent variables r, factors f, and model parameters with individual heterogeneity $\Gamma(Z)$. $\hat{\Gamma}(\cdot)$ is an estimator for $\Gamma^*(\cdot)$ via the standard empirical loss minimization criterion. For the class of **FCN** with ReLU activation, the following nonasymptotic bound holds with probability $(1 - \exp(-\gamma_1))$:

$$\mathbb{E}_T[\|\hat{\Gamma}(Z) - \Gamma^*(Z)\|_2^2] \leqslant C \left(\frac{QL \log(Q)}{T} \log T + \frac{\log\log T + \gamma_1}{T} + \epsilon_T^2 \right), \tag{6}$$

where $C > 0$ is a constant that does not depend on the number of samples T. Q and L denote the total number of parameters and network depth, respectively. **Note that this nonasymptotic bound does not necessarily hold for GCN, and the corresponding bound for GCN remains unsolved.**

3 Methodology

In this section, we first present the HeterIPCA model, then illustrate how to estimate the model parameters in the first stage. After that, we can further use the estimated parameters to infer some specific and meaningful quantities in the second stage. Our methodology has two significant advantages: nonasymptotic bound of parameter estimation and asymptotic normality of inference. Note that the former is the prerequisite of the latter.

3.1 HETEROGENEOUS IPCA

For a holistic consideration of nonlinear, interpretable, heterogeneous and SGCN-reinforced factor loading, we propose the following **HeterIPCA**:

$$r_t = \Gamma_\alpha(G_\beta(Z_t), G_\gamma(Y_t)) + G_\beta(Z_t) f_t + \epsilon_t, \tag{7}$$

where $\Gamma_\alpha \in \mathbb{R}^p$, $G_\beta \in \mathbb{R}^{p\times K}$, and $G_\gamma \in \mathbb{R}^{\mathscr{M}}$ are the intercepts (FCN), the factor loadings for the factors $f_t \in \mathbb{R}^K$ (GCN), and the α-specific instruments (GCN), respectively, which **constitute an SGCN**. $Z_t \in \mathbb{R}^{\mathscr{L}}$ and $Y_t \in \mathbb{R}^{\mathscr{M}}$ are heterogeneous instruments and public instruments, respectively. Note that for the convenience of presentation, we **align** all the relevant terms r_t, Z_t, Y_t, f_t, and ϵ_t to the **same index** t. $\Gamma_\alpha(G_\beta(Z_t), G_\gamma(Y_t))$ processes Z_t and Y_t through separate branches, then fuses them as the alpha factor (i.e., abnormal returns). This setting allows the model to learn that the impact of a global shock (in Y_t) might manifest differently depending on the position in the network and the specific attributes Z_t of an individual entity. On the other hand, the factor loadings $G_\beta(Z_t)$ explicitly map the individual-level instruments to the latent factor loading space while allowing for neighborhood influence through the graph structure. Unlike standard PCA which extracts factors from the covariance of r_t, HeterIPCA anchors the factor space to the instruments Z_t. This ensures that the factors represent the risk premiums associated specifically with the observed heterogeneous characteristics.

3.2 PARAMETER ESTIMATION

We directly denote α, β, and γ as the learnable parameters for Γ_α, G_β, and G_γ, respectively. The corresponding optimization model is:

$$\hat{\alpha}, \hat{\beta}, \hat{\gamma} = \underset{\alpha,\beta,\gamma}{\operatorname{argmin}} \mathcal{L}(\alpha, \beta, \gamma) := \frac{1}{T}\sum_{t=1}^{T} \ell(r_t, f_t, \Gamma_\alpha(G_\beta(Z_t), G_\gamma(Y_t)), G_\beta(Z_t)). \tag{8}$$

For example, the squared loss can be used for ℓ:

$$\ell(r_t, f_t, \Gamma_\alpha(G_\beta(Z_t), G_\gamma(Y_t)), G_\beta(Z_t)) = \|r_t - \Gamma_\alpha(G_\beta(Z_t), G_\gamma(Y_t)) - G_\beta(Z_t) f_t\|_2^2. \tag{9}$$

In this case, the total loss $\mathcal{L}$ is differentiable almost everywhere (abbreviated by a.e., which means excluding a zero-measure set) with respect to (w.r.t.) α, β, and γ, respectively, for the widely-used ReLU activation function. Then α, β, and γ can be learned efficiently via the backpropagation scheme of mainstream deep learning architectures (e.g., Pytorch). Specifically, the forward pass between two consecutive layers $H^{(l+1)}$ and $H^{(l)}$ for the standard Kipf & Welling architecture (Kipf & Welling, 2017) of GCN is:

$$H^{(l+1)} = \sigma(\mathcal{Z}^{(l+1)}), \quad \mathcal{Z}^{(l+1)} = \hat{A} H^{(l)} W^{(l)}, \quad \hat{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}, \tag{10}$$

where $\tilde{A} = A + \mathbb{I}$ is the adjacency matrix of an undirected graph $\mathcal{G}$ (whose structure is determined by the input data) with added self-connections $\mathbb{I}$ (whose dimensionality can be inferred by the number of nodes), and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ is the degree matrix. $W^{(l)}$ is the learnable weight matrix, and $\sigma(\cdot)$ is the element-wise activation function. The corresponding backpropagation scheme is:

$$\Delta^{(l+1)} := \frac{\partial \mathcal{L}}{\partial \mathcal{Z}^{(l+1)}}, \ \frac{\partial \mathcal{L}}{\partial W^{(l)}} = (\hat{A} H^{(l)})^\top \Delta^{(l+1)}, \ \Delta^{(l)} = (\hat{A} \Delta^{(l+1)} (W^{(l)})^\top) \odot \sigma'(\mathcal{Z}^{(l)}), \tag{11}$$

where $\odot$ denotes the element-wise product, and σ' denotes the element-wise derivative of the activation function. Because $W^{(l)}$ does not directly connect to $W^{(l-1)}$ of the previous layer and $\partial\mathcal{L}/\partial H^{(l)}$ does not include the activation information $\sigma'(\mathcal{Z}^{(l)})$, $\partial\mathcal{L}/\partial\mathcal{Z}^{(l)}$ is used as the core increment $\Delta^{(l)}$ to be backpropagated. Since $\sigma' = \operatorname{sign}$ a.e. for $\sigma = \mathrm{ReLU}$, (11) can be computed efficiently. In addition, if the GCN has bias parameters $\{b^{(l)}\}$, they can be learned via a similar backpropagation scheme. All the learnable parameters including $\{W^{(l)}, b^{(l)}\}$ constitute β or γ for G_β or G_γ. Last, the backpropagation for the FCN Γ_α is obvious, and the backpropagation for the SGCN $\Gamma_\alpha(G_\beta(\cdot), G_\gamma(\cdot))$ can be conducted by the chain rule.

3.3 INFERENCE

After the training of SGCN, the next stage is to use them to infer some specific and meaningful term μ^* in real-world scenarios, such as economic quantity, policy effect, or other value of interest:

$$\mu^* = \mathbb{E}[\mathcal{F}(Z, Y, \Gamma_\alpha^*(G_\beta^*(Z), G_\gamma^*(Y)), G_\beta^*(Z), \tilde{f})] \in \mathbb{R}^{d_\mu}, \tag{12}$$

where $\Gamma_\alpha^*(G_\beta^*(\cdot), G_\gamma^*(\cdot))$ denotes the optimal SGCN, $\tilde{f}$ denotes the target factor values to infer μ^*, and $\mathcal{F}(\cdot)$ denotes the mapping relationship from the ingredients inside the parentheses to μ^*. To

simplify notations, let $X := (Z, Y)$, $\theta := (\alpha, \beta, \gamma)$, and $\mathscr{G}_\theta(X) := (\Gamma_\alpha(G_\beta(Z), G_\gamma(Y)), G_\beta(Z))$. Denote ∂_θ and $\partial_{\theta\theta}$ as the (sub)gradient (or Jacobian) and the Hessian operator, which can be treated as **ordinary derivatives** at $\theta(X)$ with every $X = x$ in mainstream automatic differentiating schemes. Let $\partial_\theta \mathcal{F}(x, \mathscr{G}_\theta(x), \tilde{f}) \in \mathbb{R}^{d_\mu \times d_\theta}$, $\partial_\theta \ell(r, f, \mathscr{G}_\theta(x)) \in \mathbb{R}^{d_\theta}$, and $\partial_{\theta\theta} \ell(r, f, \mathscr{G}_\theta(x)) \in \mathbb{R}^{d_\theta \times d_\theta}$ be the Jacobian of $\mathcal{F}$, the gradient of ℓ, and the Hessian of ℓ, respectively. Based on the double machine learning (DML, Chernozhukov et al. 2018; 2022b;a; Farrell et al. 2025) framework, a **debiased version** of μ should be inferred:

$$\psi(r, f, \mathscr{G}_\theta(x), \Lambda(x)) = \mathcal{F}(x, \mathscr{G}_\theta(x), \tilde{f}) - \partial_\theta \mathcal{F}(x, \mathscr{G}_\theta(x), \tilde{f}) \Lambda(x)^{-1} \partial_\theta \ell(r, f, \mathscr{G}_\theta(x)), \tag{13}$$

where $\Lambda(x) := \mathbb{E}[\partial_{\theta\theta} \ell(R, F, \mathscr{G}_\theta(X)) | X = x]$ is the population Hessian of ℓ w.r.t. θ conditioned on $X = x$, which serves as a scaling matrix that reflects the curvature of ℓ w.r.t. θ at x. Then the second term of (13) can offset the biased term $\mathcal{F}(x, \mathscr{G}_\theta(x), \tilde{f})$. In this way, ψ provides a **Neyman orthogonal score** under certain conditions (see Section 4.2). Moreover, $(\psi - \mu^*)$ is exactly the **influence function** of μ^* in the field of robust statistics, which quantifies the effect of an infinitesimal contamination at the point x on μ^*. By directly calculating this influence function, we can verify Neyman orthogonality and construct an estimate $\hat{\mu}$ that follows **asymptotic normality** (see Theorem 3). Based on standard DML, we partition the data into $\mathcal{K}$ disjoint subsets $\mathcal{I}_\mathfrak{k}$ of equal size. Then

$$\hat{\mu} := \frac{1}{\mathcal{K}} \sum_{\mathfrak{k}=1}^{\mathcal{K}} \hat{\mu}_\mathfrak{k}, \quad \hat{\mu}_\mathfrak{k} = \frac{1}{|\mathcal{I}_\mathfrak{k}|} \sum_{t \in \mathcal{I}_\mathfrak{k}} \psi(r_t, f_t, \mathscr{G}_{\hat{\theta}_\mathfrak{k}}(x_t), \hat{\Lambda}_\mathfrak{k}(x_t)), \tag{14}$$

where $\mathcal{I}_\mathfrak{k}$, $\mathscr{G}_{\hat{\theta}_\mathfrak{k}}(x_t)$, and $\hat{\Lambda}_\mathfrak{k}(x_t)$ are calculated by the observations in the complementary set $\mathcal{I}_\mathfrak{k}^\complement$. Moreover, $\mathcal{I}_\mathfrak{k}^\complement$ can be further divided into two subsets separately for $\mathscr{G}_{\hat{\theta}_\mathfrak{k}}(x_t)$ and $\hat{\Lambda}_\mathfrak{k}(x_t)$ if $\Lambda(x)$ depends on $\mathscr{G}_\theta(x)$, for general theoretical soundness. The asymptotic variance of ψ can also be estimated in the same way, which will be used later to validate asymptotic normality:

$$\hat{\Psi} := \frac{1}{\mathcal{K}} \sum_{\mathfrak{k}=1}^{\mathcal{K}} \frac{1}{|\mathcal{I}_\mathfrak{k}|} \sum_{t \in \mathcal{I}_\mathfrak{k}} (\psi(r_t, f_t, \mathscr{G}_{\hat{\theta}_\mathfrak{k}}(x_t), \hat{\Lambda}_\mathfrak{k}(x_t)) - \hat{\mu}_\mathfrak{k})^2. \tag{15}$$

4 Main Theoretical Results

We establish the theoretical foundations of HeterIPCA in this section. They are divided into two main parts: first, we derive a nonasymptotic bound for the parameter functions estimated via the SGCN, which enables HeterIPCA to learn complex and heterogeneous factor loadings from finite samples. Second, we prove the asymptotic normality of the inferred term by exploiting the debiased Neyman orthogonal score in (13). These results bridge the gap between black-box parameter estimation and rigorous statistical inference in the context of HeterIPCA.

4.1 Nonasymptotic Bound of Parameter Estimation

The nonasymptotic bound indicates the approximation capabilities of a network architecture for the population optimal θ^* by the empirical estimator $\hat{\theta}$. While the nonasymptotic bound for FCN has been established, that for GCN or even the proposed SGCN in (7) remains unsolved, which requires addressing the dependency introduced by the graph structure (Lai et al., 2026c). To do this, certain regularity conditions regarding the SGCN are needed to ensure that the neighborhood influence does not lead to vanishing or exploding representations, which are commonly used in the literature (Chernozhukov et al., 2018; 2022b;a; Lin et al., 2024b;a; Farrell et al., 2025).

Assumption 1. *$\theta^*(\cdot) := \operatorname{argmin}_\theta \mathbb{E}[\ell(R, F, \mathscr{G}_\theta(X))]$ is nonparametrically identified and uniformly bounded. Besides, there exist constants c_1, c_2, and C_ℓ such that the following inequalities hold for any $\theta(x)$ and $\tilde{\theta}(x)$:*

$$|\ell(r, f, \mathscr{G}_\theta(x)) - \ell(r, f, \mathscr{G}_{\tilde{\theta}}(x))| \leqslant C_\ell \|\theta(x) - \tilde{\theta}(x)\|_2, \tag{16}$$

$$c_1 \mathbb{E}[\|\theta(X) - \theta^*(X)\|_2^2] \leqslant \mathbb{E}[\ell(R, F, \mathscr{G}_\theta(X))] - \mathbb{E}[\ell(R, F, \mathscr{G}_{\theta^*}(X))] \leqslant c_2 \mathbb{E}[\|\theta(X) - \theta^*(X)\|_2^2], \tag{17}$$

where R and F denote the random versions of the dependent variable r and the factor f, respectively.

Denote the continuously distributed elements of X by X_c with $d_c = \dim(X_c)$, and let the rest elements of X be binary random variables without loss of generality. Let $V := (R, F, X)$ be the population random variables.

Assumption 2. *(1) All the elements of V are bounded variables. (2) The support set of X_c is the compact and connected one $[-1,1]^{d_c}$. (3) $\theta_k^*(x) \in \mathcal{W}^{p,\infty}([-1,1]^{d_c})$ as functions of x_c for $k = 1, \cdots, d_\theta$, where $\mathcal{W}^{p,\infty}([-1,1]^q)$ denotes the Hölder ball:*

$$\mathcal{W}^{p,\infty}([-1,1]^q) := \{h : \max_{\mathfrak{r}\in\mathbb{N}^q, \|\mathfrak{r}\|_1 \leqslant p} \operatorname{ess\,sup}_{u\in[-1,1]^q} |D^{\mathfrak{r}} h(u)| \leqslant 1\}. \tag{18}$$

$\operatorname{ess\,sup}$ and $D^{\mathfrak{r}}$ denote the essential supremum operator and the weak derivative operator with multi-index $\mathfrak{r}$, respectively. Positive integers p and q represent the smoothness order and the argument dimension, respectively. (4) $\max_{1\leqslant k\leqslant d_\theta} \sup_x |\theta_k^*(x)| < M$ for some $M > 0$.

Definition 1 (Pseudo-dimension)**.** *Let $\mathscr{F}$ be a function class where each element h maps a domain $\mathcal{X}$ to $\mathbb{R}$, and suppose there is a subset $\mathcal{S} = \{x_1, \cdots, x_{\mathfrak{m}}\} \subseteq \mathcal{X}$. Then $\mathcal{S}$ is pseudo-shattered by $\mathscr{F}$ if there are $\mathfrak{u}_1, \cdots, \mathfrak{u}_{\mathfrak{m}} \in \mathbb{R}$ such that for every $\mathfrak{b} \in \{0,1\}^{\mathfrak{m}}$, there exists $h_{\mathfrak{b}} \in \mathscr{F}$ that satisfies $\operatorname{sign}(h_{\mathfrak{b}}(x_i) - \mathfrak{u}_i) = \mathfrak{b}_i$ for any $1 \leqslant i \leqslant \mathfrak{m}$. $\mathfrak{u} = (\mathfrak{u}_1, \cdots, \mathfrak{u}_{\mathfrak{m}})$ is said to witness this shattering.*

Moreover, if $\mathfrak{m}$ is the maximum cardinality of $\mathcal{S}$ that can be pseudo-shattered by $\mathscr{F}$, then $\mathscr{F}$ is said to have pseudo-dimension $\mathfrak{m}$. If this maximum cardinality does not exist, then the pseudo-dimension of $\mathscr{F}$ is infinite. The pseudo-dimension of $\mathscr{F}$ is denoted by $\operatorname{Pdim}(\mathscr{F})$.

Theorem 1. *The pseudo-dimension of SGCN is bounded by:*

$$\operatorname{Pdim}(\mathrm{SGCN}) \leqslant C_d Q L \log(Q\mathfrak{n}), \tag{19}$$

where C_d is a universal constant. Q, L, and $\mathfrak{n}$ denote the number of parameters, depth, and number of nodes for the SGCN.

The proof is provided in Appendix A.1.4.

Theorem 2. *Let $\{v_t\}_{t=1}^T$ be a random sample that satisfies Assumptions 1 and 2. Let $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ be the estimator defined in (8), where the SGCN $\Gamma_\alpha(G_\beta(\cdot), G_\gamma(\cdot))$ uses ReLU activation with parameter functions $\theta(x)$ bounded by $2M$, width $J \asymp T^{d_c/2(p+d_c)} \log^2 T$, and depth $L \asymp \log T$. $\asymp$ denotes the asymptotic equivalence in order. Then*

$$\begin{aligned}
\|\hat{\theta}_k - \theta_k^*\|_{L_2(X)}^2 &\leqslant \mathcal{C} \cdot \{T^{-\frac{p}{p+d_c}} \log^7 T \cdot \log(\mathfrak{n}T) + \frac{\log\log T}{T}\}, \\
\mathbb{E}_T[(\hat{\theta}_k - \theta_k^*)^2] &\leqslant \mathcal{C} \cdot \{T^{-\frac{p}{p+d_c}} \log^7 T \cdot \log(\mathfrak{n}T) + \frac{\log\log T}{T}\}
\end{aligned} \tag{20}$$

for large enough T with probability $(1 - \exp\{-T^{\frac{d_c}{p+d_c}} \log^7 T \cdot \log(\mathfrak{n}T)\})$, $k = 1, \cdots, d_\theta$. $\mathcal{C}$ is a constant that may depend on $\dim(V)$, d_θ, and other fixed terms in Assumptions 1 and 2, but not T.

The proof is provided in Appendix A.1.5, which not only **extends** previous theoretical results from FCNs to SGCNs, but also **substantially reworks** some incomplete theoretical statements in previous works regarding the use of $L_\infty(X)$ space, multi-dimensionality of θ, and Bernstein inequality.

4.2 Asymptotic Normality of Inference

To establish the asymptotic normality of the inference $\hat{\mu}$ in (14), a crucial step is to verify the Neyman orthogonality of ψ in (13). This ensures that $\hat{\mu}$ is insensitive to first-order perturbations in the SGCN parameters θ. To do this, we need further assumptions regarding sufficient regularity of ψ, which are also commonly used in the literature (Chernozhukov et al., 2018; 2022b;a; Farrell et al., 2025).

Assumption 3. *The following assumptions hold uniformly in the corresponding conditioning elements on the distribution of V: (1) $\theta^*(x)$ exists and $\ell(r, f, \mathscr{G}_\theta(x))$ is thrice continuously differentiable w.r.t. θ. (2) $\mathbb{E}[\partial_\theta \ell(R, f, \mathscr{G}_\theta(x)) | F = f, X = x] = 0$. (3) $\Lambda(x)$ has bounded inverse. (4) $\mathcal{F}(X, \mathscr{G}_\theta(X), \tilde{f})$ and $\ell(R, F, \mathscr{G}_\theta(X))$ have $q > 4$ finite absolute moments and positive variances.*

Theorem 3. *Let $\{v_t\}_{t=1}^T$ be a random sample that satisfies Assumption 3. Assume $\|\hat{\theta}_{k_1} - \theta_{k_1}^*\|_{L_2(X)} = o_{\mathfrak{p}}(T^{-\frac{1}{4}})$ and $\|(\hat{\Lambda})_{k_1,k_2} - (\Lambda)_{k_1,k_2}\|_{L_2(X)} = o_{\mathfrak{p}}(T^{-\frac{1}{4}})$ for any $k_1, k_2 \in \{1, \cdots, d_\theta\}$, and that $\hat{\Lambda}(x_t)$ is uniformly invertible. Then, (1) ψ in (13) is a Neyman orthogonal score; (2) $\hat{\mu}$ in (14) and $\hat{\Psi}$ in (15) satisfy:*

$$\sqrt{\mathcal{K}} \hat{\Psi}^{-\frac{1}{2}} (\hat{\mu} - \mu^*) = \frac{1}{\sqrt{\mathcal{K}T}} \sum_{\mathfrak{k}=1}^{\mathcal{K}} \Psi^{-\frac{1}{2}} \psi(r_{\mathfrak{k}}, f_{\mathfrak{k}}, \mathscr{G}_{\theta^*}(x_{\mathfrak{k}}), \Lambda(x_{\mathfrak{k}})) + o_{\mathfrak{p}}(1) \xrightarrow{d} \mathcal{N}(0_{d_\mu}, \mathbb{I}_{d_\mu}) \tag{21}$$

where Ψ denotes the ground-truth version of $\hat{\Psi}$, $o_{\mathfrak{p}}$ denotes an infinitesimal term in probability, $\xrightarrow{d}$ denotes the convergence in distribution, and $\mathcal{N}(0_{d_\mu}, \mathbb{I}_{d_\mu})$ denotes the d_μ-dimensional standard normal distribution.

The proof is provided in Appendix A.2. Theorem 3 provides a rigorous foundation for statistical inference on the specific quantity μ. It is worth noting that the asymptotic variance $\hat{\Psi}$ accounts for the cross-sectional heterogeneity through the SGCN $\mathscr{G}_{\hat{\theta}_{\mathfrak{k}}}$. By incorporating this idiosyncratic volatility, the inference of HeterIPCA achieves better efficiency than standard IPCA in settings where variances are highly dispersed. This ensures that the resulting factor exposures are not unduly influenced by high-noise variables, which leads to more stable and interpretable risk premiums in empirical applications.

4.3 Interpretability of HeterIPCA

Theorems 2 and 3 exactly solve the **factor identification** and **statistical interpretability** problems by exploiting the nonlinear factor loadings $G_\beta(Z_t)$. In IPCA, the instruments Z_t are **observable** variables, which are identifiable. Theorem 2 establishes a nonasymptotic error bound for parameter estimation, which ensures that $G_\beta(Z_t)$ is **anchored** to Z_t. Hence $G_\beta(Z_t)$ is also identifiable. Then the factors f_t can be **determined** by $G_\beta(Z_t)$ and thus identifiable. In this case, there is no need to consider other rotations $(G_\beta(Z_t)\mathcal{R}^{-1})(\mathcal{R}f_t)$. Because $G_\beta(Z_t)$ is already a **meaningful nonlinear mapping** of Z_t, it is redundant and inappropriate to add another linear rotation mapping $\mathcal{R}^{-1}$. Moreover, Theorem 3 further ensures that the **inference on identified factors** achieves **asymptotic normality**, and the **identified factor exposures** remain statistically **stable and interpretable** in volatile environments (the asymptotic variance term $\hat{\Psi}$).

In contrast, standard IPCA uses a linear factor loading form $Z_t\Gamma_\beta f_t$, which has **arbitrary linear rotational variants** $Z_t(\Gamma_\beta\mathcal{R}^{-1})(\mathcal{R}f_t)$. Since $\Gamma_\beta\mathcal{R}^{-1}$ is also a linear transform, it is difficult to tell whether it is better than the linear transform Γ_β or not. Hence the ambiguity of factorization exists in standard IPCA. As for other baselines, CA considers neither heterogeneity nor abnormal factor. FactorGCL drops both the form of IPCA and the consistency of statistical inference. These three baselines do not have theoretical supports in parameter estimations and statistical inference. Hence they are less interpretable than our method. More experimental results regarding interpretability are provided in Appendix A.9.

5 Experiment

In this section, we conduct experiments to evaluate HeterIPCA in terms of predictive performance, factor interpretability, and its ability to capture heterogeneous instrument-loading relationships. The experimental setup and implementation details are provided in Appendices A.3 and A.4, respectively. In brief, we follow most of the standard conventions in the asset pricing literature (Gu et al., 2020; 2021; Kelly et al., 2019) with multiple continuous factors, which represents a certain difficulty in the asset pricing practice. We further provide experiments on **statistical significance, turnover analysis, SHAP analysis, ablation analysis, interpretability analysis, number of latent factors** K in **Appendices A.5∼A.10**, which strengthen the main conclusions of HeterIPCA.

Baselines. To keep economic meaning, only baselines that preserve the standard decomposition of returns into idiosyncratic mispricing (α) and systematic risk ($\beta \cdot f$, $\Gamma_\beta \cdot f$, or $\Gamma_\beta(f)$) are relevant for comparison. Specifically, we consider six strong, widely-used, and structured baselines in empirical asset pricing: PCA, IPCA, CA (four different versions denoted by CA0∼CA3), SDNN, FactorGCL, and CVAB (Liu et al., 2025).

Evaluating Metrics. The first metric is the **out-of-sample Predictive** R^2 (109), which measures the forward-forecasting capability of a model and the effectiveness of **parameter estimation**. The second metric is the **out-of-sample Sharpe ratio (SR)** defined in (110) (Sharpe, 1966), which measures the risk-adjusted return in a real-world investing task with transaction costs and evaluates the performance of **statistical inference**. Hence these two metrics correspond to the two-stage implementation of HeterIPCA in Sections 3.2 and 3.3.

5.1 Predictive Performance (Parameter Estimation)

Table 1 summarizes the predictive performance of HeterIPCA and baselines with $K = 1 \sim 6$ latent factors, evaluated at both the stock-level R^2_r and the characteristic-managed portfolio-level R^2_{x}. Similar to the findings in recent asset pricing literature, both R^2_r and R^2_{x} generally increase

Table 1: Predictive performance on the U.S. equity data set. R^2_r evaluates the stock-level explanatory power, while R^2_{x} evaluates the portfolio-level explanatory power. Higher values indicate better performance. The values in bold (underlined) indicate the best (second best) results.

Model	R^2_r (%)						R^2_{x} (%)					
	K=1	K=2	K=3	K=4	K=5	K=6	K=1	K=2	K=3	K=4	K=5	K=6
PCA	0.207	0.218	0.216	0.218	0.219	0.218	0.352	0.392	0.396	0.405	0.408	0.410
IPCA	0.069	-2.977	0.275	-0.163	0.364	<u>0.651</u>	0.462	0.317	0.427	0.418	0.429	0.428
CA0	0.159	0.214	0.241	0.259	0.293	0.266	<u>0.558</u>	<u>0.666</u>	0.229	0.443	0.574	0.429
CA1	0.147	0.149	0.156	0.161	0.167	0.185	0.490	0.473	0.552	0.586	0.643	0.735
CA2	0.165	0.148	0.163	0.161	0.166	0.159	**0.638**	0.500	0.645	0.588	0.607	0.582
CA3	0.139	0.143	0.156	0.163	0.149	0.148	0.516	0.508	0.631	<u>0.688</u>	0.590	0.551
SDNN	0.215	0.192	0.140	0.071	0.096	0.147	0.232	-0.353	-0.510	-1.124	-0.90	-0.748
FactorGCL	<u>0.286</u>	<u>0.351</u>	<u>0.338</u>	<u>0.412</u>	<u>0.439</u>	0.304	0.503	0.271	<u>0.686</u>	0.677	<u>1.235</u>	<u>1.127</u>
CVAB	0.264	0.335	0.312	0.395	0.418	0.288	0.461	0.254	0.652	0.649	1.198	1.085
HeterIPCA (Ours)	**0.482**	**0.558**	**0.562**	**0.643**	**0.646**	**0.664**	0.473	**0.931**	**1.095**	**1.318**	**1.583**	**1.560**

along with K, as a larger K accommodates more information. However, for most baseline models, the performance tends to reach its peak early or even deteriorate rapidly as K increases. This phenomenon reveals a severe problem regarding the trade-off between information acquisition and model overfitting. Compared with these baselines, HeterIPCA achieves the best predictive performance by significantly large margins across almost all K configurations. For the stock-level R^2_r, HeterIPCA shows robust scalability and reaches an R^2_r of 0.664% at $K = 6$. In contrast, while advanced representation learning baselines like FactorGCL and CVAB show competitive performance at lower dimensions, their R^2_r values drop significantly to 0.304% and 0.288% at $K = 6$, respectively. Further analysis in Appendix A.5 indicates that HeterIPCA achieves **superior performance** than its three primary ablation configurations across different K with **high statistical significance**.

For the portfolio-level R^2_{x}, the advantage of HeterIPCA over other baselines is even more significant. It achieves a peak R^2_{x} of 1.583% at $K = 5$, which exceeds the R^2_{x} of the best-performing baseline FactorGCL (1.235%) by a considerable margin of 28.18%, and even more significantly outperforms the traditional IPCA (0.429%). These results indicate that HeterIPCA possesses far more robust predictive power for future returns, and efficiently translates latent representations into generalizable economic signals. It is worth noting that while FactorGCL and CVAB incorporate sophisticated modeling techniques, their R^2_{x} values still drop to 1.127% and 1.085% at $K = 6$, which shows a susceptibility to historical market noise at a higher latent dimensionality. Besides, the R^2_{x} of SDNN collapses to only -0.748% at $K = 6$, which indicates that SDNN may have learned a large proportion of historical market noise. Hence a structured architecture does not necessarily lead to an effective representation, and the FCN-based SDNN should be improved to the GCN-based HeterIPCA. Specifically, HeterIPCA maintains the notably highest R^2_{x} of 1.560% in this case, which indicates that the proposed SGCN serves as an effective structural regularizer. By modeling the cross-asset spillover effects and inter-dependencies, HeterIPCA avoids memorizing idiosyncratic noise and captures the true common risk factors that drive cross-sectional return variations.

5.2 Economic Evaluation (Inference)

To evaluate the economic significance and practical viability of HeterIPCA, we conduct an empirical out-of-sample portfolio backtest based on the decile ranking of predicted asset returns. We perform a comparative evaluation of all candidate models across varying latent factor counts $K \in \{1, 2, \ldots, 6\}$. Specifically, at the end of each month t, assets are sorted by their predicted returns $\hat{r}_{i,t+1}$ to identify the top and bottom 10% of the cross-section. We construct two types of portfolios: (1) Long-only portfolios: comprising assets in the top decile, rebalanced monthly; (2) Long-short zero-investment portfolios: formed by taking long positions in the top decile and short positions in the bottom decile, with the funding leg matched to the size of the investment leg. All portfolios use equal weighting and monthly rebalancing. To incorporate realistic market frictions and account for turnover, we impose a conservative transaction cost of 30 basis points (bps) per rebalancing event, which is a stringent threshold designed to penalize excessive trading. Finally, economic performance is quantified by the annualized SRs of the constructed portfolios on the out-of-sample test data.

Table 2 summarizes the out-of-sample SRs for both long-only and long-short portfolios across all evaluated baselines. HeterIPCA consistently and significantly outperforms all these baselines

Table 2: Out-of-sample Sharpe ratios of long-short and long-only portfolios with 30 bps transaction costs across different numbers of latent factors (K). The values in bold (underlined) indicate the best (second best) results.

Model	Long-short Portfolios						Long-only Portfolios					
	K=1	K=2	K=3	K=4	K=5	K=6	K=1	K=2	K=3	K=4	K=5	K=6
PCA	-0.45	-0.52	-0.46	-0.38	-0.39	0.12	0.08	0.21	-0.44	0.24	0.27	0.24
IPCA	0.26	0.13	0.11	0.87	1.22	1.28	0.50	0.39	0.48	0.88	0.89	0.87
CA0	0.30	0.51	1.27	1.64	1.43	1.63	0.45	0.61	0.89	<u>0.93</u>	0.87	0.91
CA1	0.27	0.32	0.33	0.38	0.39	0.39	0.41	0.48	0.50	0.54	0.55	0.56
CA2	0.29	0.32	0.37	0.36	0.38	0.39	0.44	0.46	0.50	0.51	0.52	0.53
CA3	0.26	0.29	0.31	0.28	0.29	0.28	0.44	0.46	0.42	0.44	0.45	0.44
SDNN	<u>1.48</u>	1.54	1.64	1.54	1.51	1.36	0.78	0.82	0.81	0.87	0.81	0.82
FactorGCL	0.62	<u>2.14</u>	<u>2.07</u>	<u>2.06</u>	<u>2.12</u>	<u>2.05</u>	0.58	<u>0.94</u>	0.91	<u>0.93</u>	<u>0.97</u>	0.90
CVAB	0.97	1.09	1.53	1.49	1.78	1.80	<u>0.88</u>	0.85	<u>0.94</u>	0.83	<u>0.97</u>	<u>0.99</u>
HeterIPCA (Ours)	**2.39**	**2.33**	**2.52**	**2.44**	**2.55**	**2.57**	**1.05**	**1.06**	**1.07**	**1.04**	**1.09**	**1.12**

across all K configurations on this risk-adjusted performance. As expected, the long-short strategies generally yield higher Sharpe ratios than their long-only counterparts for the top-performing models, which validates their capacity to identify underperforming assets and the subsequent benefits of portfolio hedging. While traditional linear factor models (such as PCA) fail to maintain profitability under the 30 bps transaction cost threshold, advanced representation baselines (such as FactorGCL and CVAB) show greater resilience and emerge as the strongest competitors.

Nevertheless, HeterIPCA maintains a substantially superior economic profile. For $K = 6$, the long-short SR of 2.57 for HeterIPCA represents a 25.37% improvement over FactorGCL (2.05), a 57.67% improvement over CA0 (1.63), and an 88.97% improvement over SDNN (1.36). Similarly, in the long-only setting for $K = 6$, HeterIPCA achieves an SR of 1.12, which surpasses CVAB (0.99) by 13.13%. We emphasize that this superior net profitability is not a consequence of reduced trading frequency; the monthly turnover rates for HeterIPCA are highly comparable to other deep learning baselines, as detailed in Appendix A.6. Instead, the performance gain is driven by superior gross returns, which reflects more precise predictive rankings. We attribute this success to the architecture design of HeterIPCA. While FactorGCL and CVAB incorporate advanced structural or variational techniques, they often struggle to disentangle systemic risk exposures from firm-specific mispricing. Besides, while the FCN architecture of SDNN evaluates assets separately and independently, the SGCN architecture of HeterIPCA captures topological dependencies and cross-asset spillover effects. This structural inductive bias facilitates a more accurate sorting of cross-sectional anomalies, which enables HeterIPCA to effectively absorb significant transaction costs while yielding robust economic returns.

6 Conclusion

Instrumented principal component analysis (IPCA) is an emerging approach that simultaneously estimates the latent factors and recovers their corresponding loadings. However, it can support neither individual-level heterogeneity nor non-linear instrumental representation. To address these problems, we propose an interpretable Heterogeneous Instrumented Principal Component Analysis (HeterIPCA), which incorporates a Structured Graph Convolutional Network (SGCN). HeterIPCA integrates micro-level heterogeneous instruments (through a graph structure) and macro-level public instruments to capture both global market trends and local context-dependent dependencies. It flexibly models factor loadings and abnormal returns while preserving essential econometric interpretability. Moreover, we establish a rigorous mathematical foundation for this deep learning-based approach, including: (1) a sharp nonasymptotic bound for parameter estimation by calculating the pseudo-dimension of the SGCN; (2) the asymptotic normality of inference by exploiting debiased Neyman orthogonal scores from the double machine learning approach. This contribution bridges the critical gap between high-capacity black-box models and formal statistical inference.

Experimental results show that HeterIPCA achieves effective performance across multiple evaluation metrics compared with both linear and non-linear state-of-the-art baselines. These findings offer new insights into the interpretability and heterogeneity of factor analysis, and also provide a robust and theoretically sound approach for analyzing high-dimensional data with complex entity relationships.

REFERENCES

Donald WK Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica: Journal of the Econometric Society*, pp. 817–858, 1991.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, USA, 1st edition, 2009. ISBN 052111862X.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Management Science*, 70(2):714–750, 2024.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022a.

Victor Chernozhukov, Whitney Newey, Víctor M Quintas-Martínez, and Vasilis Syrgkanis. RieszNet and ForestRiesz: Automatic debiased machine learning with neural nets and random forests. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3901–3914. PMLR, 17–23 Jul 2022b.

Carter Davis, Mahyar Kargar, and Jiacui Li. Why do portfolio choice models predict inelastic demand? *Journal of Financial Economics*, 172:104096, 2025. ISSN 0304-405X.

Yitong Duan, Weiran Wang, and Jian Li. Factorgcl: A hypergraph-based factor model with temporal residual contrastive learning for stock returns prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1):173–181, Apr. 2025.

Eugene F Fama and Kenneth R French. Industry costs of equity. *Journal of financial economics*, 43 (2):153–193, 1997.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep learning for individual heterogeneity: An automatic inference framework. *arXiv preprint arXiv:2010.14694v3*, 2025.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 02 2020.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Autoencoder asset pricing models. *Journal of Econometrics*, 222(1, Part B):429–450, 2021. ISSN 0304-4076. Annals Issue: Financial Econometrics in the Age of the Digital Economy.

Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In Satyen Kale and Ohad Shamir (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1064–1068. PMLR, 07–10 Jul 2017.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Bryan T. Kelly, Seth Pruitt, and Yinan Su. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524, 2019. ISSN 0304-405X.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Zhao-Rong Lai and Weiwen Wang. Invariant risk minimization is a total variation model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 25913–25935. PMLR, 21–27 Jul 2024.

Zhao-Rong Lai and Haisheng Yang. Linear trading position with sparse spectrum. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 5554–5562. International Joint Conferences on Artificial Intelligence Organization, 8 2025. Main Track.

Zhao-Rong Lai, Xiwen Yuan, Ziliang Chen, Liangda Fang, and Yongsen Zheng. A unified total variation framework for membrane potential perturbation dynamic. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Zhao-Rong Lai, Xiwen Yuan, and Jian Weng. Adversarial attack and defense for denoising diffusion sampling. In *The 43rd International Conference on Machine Learning*, 2026b.

Zhao-Rong Lai, Zheng-Sen Zhou, Liangda Fang, Yongsen Zheng, and Ziliang Chen. ℓ_1 latent distance based continuous-time graph representation. In *The Fourteenth International Conference on Learning Representations*, 2026c.

Markus Leippold, Qian Wang, and Wenyu Zhou. Machine learning in the chinese stock market. *Journal of Financial Economics*, 145(2, Part A):64–82, 2022. ISSN 0304-405X.

Yizun Lin, Zhao-Rong Lai, and Cheng Li. A globally optimal portfolio for m-sparse sharpe ratio maximization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 17133–17160, 2024a.

Yizun Lin, Yangyu Zhang, Zhao-Rong Lai, and Cheng Li. Autonomous sparse mean-CVaR portfolio optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30440–30456. PMLR, 21–27 Jul 2024b.

Ruirui Liu, Huichou Huang, and Johannes Ruf. Asset pricing with contrastive adversarial variational bayes. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 7572–7579, 2025.

Shahar Mendelson. *A Few Notes on Statistical Learning Theory*, pp. 1–40. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

William F. Sharpe. Mutual fund performance. *Journal of Business*, 39(1):119–138, Jan. 1966.

Namrata Vaswani, Yuejie Chi, and Thierry Bouwmans. Rethinking PCA for modern data sets: Theory, algorithms, and applications. *Proceedings of the IEEE*, 106(8):1274–1276, 2018.

Yuanchao Wang, Zhao-Rong Lai, and Tianqi Zhong. Out-of-distribution generalization for total variation based invariant risk minimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Hugh E. Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968.

Ivo Welch and Amit Goyal. A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508, 2008.

Quanzhou Xiang, Zhan Chen, Qi Sun, and Rujun Jiang. Rsap-dfm: Regime-shifting adaptive posterior dynamic factor model for stock returns prediction. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. International Joint Conferences on Artificial Intelligence Organization*, 2024.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017. ISSN 0893-6080.

A Appendix

A.1 Proof of Theorems 1 and 2

Before digging into the main proofs of Theorems 1 and 2, we provide some preliminaries regarding Bernstein inequality, Rademacher complexity, pseudo-dimension, and localization.

A.1.1 Bernstein Inequality

Lemma A1 (Bernstein Inequality). *Let $\{\xi_t\}_{t=1}^T$ be independent mean-zero random variables. Suppose $|\xi_t| \leqslant \mathscr{M}$ a.e. for all t. Then for all $\delta > 0$,*

$$\mathbb{P}\{(\sum_{t=1}^{T} \xi_t) \geqslant \delta\} \leqslant \exp\left(-\frac{\delta^2/2}{\nu_T^2 + \mathscr{M}\delta/3}\right), \tag{22}$$

where $\nu_T^2 := \sum_{t=1}^T \mathbb{E}[\xi_t^2]$.

Lemma A2. *Using the same notations and conditions as those of Lemma A1, the following inequality holds with probability at least $(1 - \exp(-\zeta))$ ($\zeta > 0$):*

$$\sum_{t=1}^{T} \xi_t < \sqrt{2\zeta}\nu_T + \frac{2\zeta\mathscr{M}}{3}. \tag{23}$$

Proof of Lemma A2. We let $\zeta := \frac{\delta^2/2}{\nu_T^2 + \mathscr{M}\delta/3}$ in (22), then solve for δ inversely from ζ. Taking the larger root of δ (to ensure $\delta > 0$), we have

$$\begin{aligned}
\delta =& \frac{\zeta\mathscr{M}}{3} + \sqrt{\frac{\zeta^2\mathscr{M}^2}{9} + 2\zeta\nu_T^2} \\
\leqslant& \frac{\zeta\mathscr{M}}{3} + \sqrt{\frac{\zeta^2\mathscr{M}^2}{9} + 2 \cdot \frac{\zeta\mathscr{M}}{3} \cdot \sqrt{2\zeta}\nu_T + 2\zeta\nu_T^2} \\
\leqslant& \frac{\zeta\mathscr{M}}{3} + \left(\frac{\zeta\mathscr{M}}{3} + \sqrt{2\zeta}\nu_T\right) \\
=& \frac{2\zeta\mathscr{M}}{3} + \sqrt{2\zeta}\nu_T,
\end{aligned} \tag{24}$$

where the first inequality holds by adding a nonnegative term. Inserting (24) into (22), we derive (23).

□

A.1.2 Rademacher Complexity

Definition A1 (Rademacher Complexity). *Let $\varsigma_t = \pm 1$ be independent and identically distributed (i.i.d.) Rademacher random variables with $\mathbb{P}\{\varsigma_t = 1\} = \mathbb{P}\{\varsigma_t = -1\} = 1/2$. Define the following random variable of a function h or a function class $\mathscr{F}$:*

$$\mathcal{R}_T h := \frac{1}{T}\sum_{t=1}^{T} \varsigma_t h(X_t), \quad \mathcal{R}_T\mathscr{F} := \sup_{h\in\mathscr{F}} \mathcal{R}_T h. \tag{25}$$

It indicates the flexibility of h or $\mathscr{F}$ in random sign prediction. Then the (expected) Rademacher complexity is $\mathbb{E}\mathcal{R}_T\mathscr{F}$, and the empirical Rademacher complexity is defined as:

$$\mathbb{E}_\varsigma\mathcal{R}_T\mathscr{F} := \frac{1}{T}\mathbb{E}[\sup_{h\in\mathscr{F}} \sum_{t=1}^{T} \varsigma_t h(X_t)|X_1, \cdots, X_T]. \tag{26}$$

Definition A2 (Metric Entropy). *Let $\mathscr{F}$ be a function class of a function space $\mathscr{X}$ with a well-defined metric $\mathfrak{d}$. Then the ρ-covering number of $(\mathscr{F}, \mathfrak{d})$ is defined by the smallest number $\mathcal{E}(\rho, \mathscr{F}, \mathfrak{d})$ such that:*

$$\mathscr{F} \subseteq \bigcup_{i=1}^{\mathcal{E}(\rho,\mathscr{F},\mathfrak{d})} \mathcal{B}_\mathfrak{d}(h_i, \rho), \quad \{h_i\}_{i=1}^{\mathcal{E}(\rho,\mathscr{F},\mathfrak{d})} \subseteq \mathscr{F}. \tag{27}$$

where $\mathcal{B}_{\mathfrak{d}}(h_i, \rho)$ denotes the closed ball centered at h_i with radius ρ. The metric entropy is defined by $\log \mathcal{E}(\rho, \mathscr{F}, \mathfrak{d})$.

Next, we recall some important properties of Rademacher complexity in the following lemma.

Lemma A3 (Theorem 2.25 in (Mendelson, 2003) and Lemma 3 in (Farrell et al., 2021))**.** *(1) If $\phi : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with constant C_ϕ and satisfies $\phi(0) = 0$, then*

$$\mathbb{E}_\varsigma \mathcal{R}_T\{\phi \circ h : h \in \mathscr{F}\} \leqslant 2C_\phi \mathbb{E}_\varsigma \mathcal{R}_T \mathscr{F}. \tag{28}$$

(2) Let $\|h\|_T := (\mathbb{E}_T[h(x_t)^2])^{1/2}$ be the empirical function-L_2-norm of h on samples $\{x_t\}_{t=1}^T$, and $\log \mathcal{E}(\rho, \mathscr{F}, \|\cdot\|_T)$ be the metric entropy for $\mathscr{F}$ that covers radius ρ under metric $\|\cdot\|_T$. Then

$$\mathbb{E}_\varsigma \mathcal{R}_T\{h : h \in \mathscr{F}, \|h\|_T \leqslant \iota\} \leqslant \inf_{0 \leqslant \varrho \leqslant \iota} \left\{ 4\varrho + \frac{12}{\sqrt{T}} \int_\varrho^\iota \sqrt{\log \mathcal{E}(\rho, \mathscr{F}, \|\cdot\|_T)}\, \mathrm{d}\rho \right\}. \tag{29}$$

Moreover, $\mathcal{E}(\rho, \mathscr{F}, \|\cdot\|_T) \leqslant \mathcal{E}(\rho, \mathscr{F}|_{\{x_t\}_{t=1}^T}, L_\infty)$ because $\|h\|_T \leqslant \max_t |h(x_t)|$. Hence (29) also holds with $\mathcal{E}(\rho, \mathscr{F}, \|\cdot\|_T)$ substituted by $\mathcal{E}(\rho, \mathscr{F}|_{\{x_t\}_{t=1}^T}, L_\infty)$, where $\mathscr{F}|_{\{x_t\}_{t=1}^T}$ denotes the projection of the function class $\mathscr{F}$ onto the given data.

Lemma A3 indicates that the Rademacher complexity can be bounded by the metric entropy. The following lemma further indicates that the metric entropy can be bounded by the pseudo-dimension (see Definition 1), which is closely related to the Vapnik-Chervonenkis (VC) dimension.

Lemma A4 (Theorems 12.2 and 14.1 in (Anthony & Bartlett, 2009))**.** *Denote the pseudo-dimension of $\mathscr{F}$ as $\mathrm{Pdim}(\mathscr{F})$, and assume $\|h\|_{L_\infty(X)} \leqslant \mathscr{M}$. Then for $T \geqslant \mathrm{Pdim}(\mathscr{F})$ and any ρ,*

$$\mathcal{E}(\rho, \mathscr{F}|_{\{x_t\}_{t=1}^T}, L_\infty) \leqslant \left(\frac{2e\mathscr{M}T}{\rho \mathrm{Pdim}(\mathscr{F})} \right)^{\mathrm{Pdim}(\mathscr{F})}. \tag{30}$$

Using the above properties of Rademacher complexity, the empirical process terms in part ② of (47) can be bounded by the following type of Talagrand's concentration inequality.

Lemma A5 (Theorem 2.1 in (Bartlett et al., 2005) and Lemma 5 in (Farrell et al., 2021))**.** *For any g in a function class $\mathscr{H}$, assume $|g| \leqslant \mathcal{U}$ and $\mathbb{V}(g) \leqslant \mathcal{V}$, where $\mathbb{V}(g)$ denotes the variance of g. Then for any ζ, with probability at least $(1 - \exp(-\zeta))$,*

$$\sup_{g \in \mathscr{H}} \{\mathbb{E}g - \mathbb{E}_T g\} \leqslant 3\mathbb{E}\mathcal{R}_T \mathscr{H} + \sqrt{\frac{2\mathcal{V}\zeta}{T}} + \frac{4\mathcal{U}\zeta}{3T}, \tag{31}$$

and with probability at least $(1 - 2\exp(-\zeta))$,

$$\sup_{g \in \mathscr{H}} \{\mathbb{E}g - \mathbb{E}_T g\} \leqslant 6\mathbb{E}_\varsigma \mathcal{R}_T \mathscr{H} + \sqrt{\frac{2\mathcal{V}\zeta}{T}} + \frac{23\mathcal{U}\zeta}{3T}. \tag{32}$$

The same result holds for $\sup_{g \in \mathscr{H}} \{\mathbb{E}_T g - \mathbb{E}g\}$.

The following lemma embeds a DNN architecture into the FCN architecture while maintaining a scalable approximation property.

Lemma A6. ***(1) Theorems 1 and 3 in (Harvey et al., 2017).*** *Consider a DNN architecture $\mathscr{F}_{DNN} := \mathscr{F}_{DNN}(Q, L)$ with ReLU activation functions, number of parameters Q, and depth L. Then its pseudo-dimension satisfies*

$$c_f QL \log(Q/L) \leqslant \mathrm{Pdim}(\mathscr{F}_{DNN}) \leqslant C_f QL \log(Q), \tag{33}$$

where $c_f, C_f > 0$ are universal constants.

(2) Lemma 1 in (Farrell et al., 2021). *For any $\mathfrak{f} \in \mathscr{F}_{DNN}$, there exists a $\mathfrak{g} \in \mathscr{F}_{FCN}$ with $J \leqslant QL + \mathscr{U}$ such that $\mathfrak{g} = \mathfrak{f}$, where $\mathscr{U}$ denotes the total number of units in the FCN.*

(3) Theorem 6 in (Yarotsky, 2017). *$\mathscr{F}_{DNN}$ approximates the Hölder ball $\mathcal{W}^{p,\infty}([-1,1]^q)$ in the following sense: for any $\mathfrak{f}^* \in \mathcal{W}^{p,\infty}([-1,1]^q)$, there exists a $\mathfrak{f}_T(\varepsilon) \in \mathscr{F}_{DNN}$ such that*

$$\|\mathfrak{f}_T - \mathfrak{f}^*\|_{L_\infty(X)}, L(\varepsilon) \leqslant \mathcal{C}(\log(\frac{1}{\varepsilon}) + 1), \max\{Q(\varepsilon), \mathscr{U}(\varepsilon)\} \leqslant \mathcal{C}\varepsilon^{-\frac{q}{p}}(\log(\frac{1}{\varepsilon}) + 1), \tag{34}$$

where $\mathcal{C}$ only depends on p and q.

A.1.3 LOCALIZATION METHOD

The localization method (Bartlett et al., 2005; Farrell et al., 2021) aims to improve the conservative generalization upper bound of the global Rademacher Complexity towards a more realistic and tightened bound via the empirical process, which is restricted to a specific variance-based neighborhood.

Lemma A7 (Quadratic Process, Farrell et al. 2021)**.** *Use the same assumptions as in Lemma A5. If* $\|\theta_k - \theta_k^*\|_{L_2(X)} \leqslant \iota$*, then* $\|\theta_k - \theta_k^*\|_T \leqslant 2\iota$ *with probability at least* $(1 - \exp(-\zeta))$ *,* $k = 1, \cdots, d_\theta$.

Proof. Let $g_k = (\theta_k - \theta_k^*)^2$, then

$$|g_k| \leqslant (2M + M)^2 = (3M)^2, \tag{35}$$

$$\mathbb{V}(g_k) \leqslant \mathbb{E}[g_k^2] \leqslant \mathbb{E}[(\theta_k - \theta_k^*)^4] \leqslant \mathbb{E}[(|\theta_k| + |\theta_k^*|)^2(\theta_k - \theta_k^*)^2] \leqslant (3M)^2\iota^2. \tag{36}$$

Applying Lemma A5 to g_k and Lemma A3 (1) to $(\theta_k - \theta_k^*)^2$, with probability at least $(1 - \exp(-\zeta))$,

$$\begin{aligned}
&\mathbb{E}_T[g_k] - \mathbb{E}[g_k] \\
\leqslant &3\mathbb{E}\mathcal{R}_T\{(\theta_k - \theta_k^*)^2 : \mathscr{G}_\theta \in \mathrm{SGCN}, \|\theta_k - \theta_k^*\|_{L_2(X)} \leqslant \iota\} + 3M\iota\sqrt{\frac{2\zeta}{T}} + \frac{36M^2\zeta}{3T} \\
\leqslant &18M\mathbb{E}\mathcal{R}_T\{\theta_k - \theta_k^* : \mathscr{G}_\theta \in \mathrm{SGCN}, \|\theta_k - \theta_k^*\|_{L_2(X)} \leqslant \iota\} + 3M\iota\sqrt{\frac{2\zeta}{T}} + \frac{12M^2\zeta}{T},
\end{aligned} \tag{37}$$

where the second inequality holds because $(\theta_k - \theta_k^*)^2 \leqslant 3M|\theta_k - \theta_k^*|$.

Let the radius ι satisfies

$$\iota^2 \geqslant 18M\mathbb{E}\mathcal{R}_T\{\theta_k - \theta_k^* : \mathscr{G}_\theta \in \mathrm{SGCN}, \|\theta_k - \theta_k^*\|_{L_2(X)} \leqslant \iota, \forall k\}, \tag{38}$$

$$\iota^2 \geqslant \frac{6\sqrt{6}M^2\zeta}{T}. \tag{39}$$

Then (37) leads to

$$\mathbb{E}_T[(\theta_k - \theta_k^*)^2] \leqslant \iota^2 + \iota^2 + 3M\iota\sqrt{\frac{2\zeta}{T}} + \frac{12M^2\zeta}{T} \leqslant (2\iota)^2, \forall k, \tag{40}$$

which finishes the proof.

□

A.1.4 PROOF OF THEOREM 1

Proof. Recall the pseudo-dimension of FCN in Lemma A6 and the architecture of SGCN in (10) and (12). The essential extensions from FCN to SGCN lie in two parts: the use of **convolution operator** and the formulation of **graph structure**. Accordingly, we exploit the following two key properties to establish the pseudo-dimension of SGCN:

1. The convolution operator $W^{(l)}$ is a linear operator implemented on the current layer input $H^{(l)}$, which is essentially the same as the linear map of FCN.

2. The graph structure can also be formulated as a linear operator $\hat{A}$ implemented on $H^{(l)}$, but its complexity is related to the number of nodes $\mathfrak{n}$.

Since both $W^{(l)}$ and $\hat{A}$ are **linear operators**, we can retrieve and extend the key techniques and mechanisms in deducing the pseudo-dimension of FCN.

First, the **fundamental principle** of establishing the pseudo-dimension lies in the **piece-wise linearity** of ReLU:

1. Each ReLU output has two states: activation (> 0) and inhibition ($\leqslant 0$).

2. Given $\mathfrak{m}$ samples, the total number of decision hyperplanes is $\mathfrak{m} \cdot \mathscr{U}$, where $\mathscr{U}$ denotes the total number of units in the network.

3. The output of the network $\mathscr{G}_\theta$ is a polynomial function of the parameters θ. Since $\hat{A}$ is a linear operator, it does not change the degree of such polynomial function.

To simplify the tedious and abstruse deductions, we start with a tractable and important checkpoint from the Warren's Theorem (Warren, 1968; Harvey et al., 2017):

$$2^{\mathfrak{m}} \leqslant \prod_{l=1}^{L} 2\left(\frac{2e \cdot \mathfrak{m} \cdot \mathscr{U}_l \cdot (1+(l-1)\mathrm{Deg}^{l-1})}{Q_{(l)}}\right)^{Q_{(l)}}, \tag{41}$$

where e is the Euler's number, $\mathscr{U}_l$ denotes the number of units at layer l, Deg is the degree of the activation function w.r.t. θ, and $Q_{(l)}$ denotes the **total number of parameters** at the inputs to units in all the layers **up to layer** l. In the ReLU case, $\mathrm{Deg} = 1$. (41) indicates the upper bound for the number of combinatorial patterns of $\mathfrak{m}$ samples that can be represented by the function class of FCN with $\{\mathscr{U}_l, Q_{(l)}\}_{l=1}^{L}$.

Due to the topological structure of an SGCN, the adjacency matrix $\hat{A}$ aggregates features from the local neighborhood of each node. Therefore, the forward propagation of each layer l can involve the activations of all the $\mathfrak{n}$ nodes at most, which leads to $\mathfrak{n} \cdot \mathscr{U}_l$ piecewise thresholding operations in total. Replacing $\mathscr{U}_l$ by $\mathfrak{n} \cdot \mathscr{U}_l$ in (41) and letting $\mathrm{Deg} = 1$, we have

$$2^{\mathfrak{m}} \leqslant \prod_{l=1}^{L} 2\left(\frac{2e \cdot \mathfrak{m} \cdot \mathfrak{n} \cdot \mathscr{U}_l \cdot l}{Q_{(l)}}\right)^{Q_{(l)}} \tag{42}$$

$$\leqslant 2^L \left(\frac{2e \cdot \mathfrak{m} \cdot \mathfrak{n} \cdot (\sum_{l=1}^{L} \mathscr{U}_l \cdot l)}{\sum_{l=1}^{L} Q_{(l)}}\right)^{\sum_{l=1}^{L} Q_{(l)}} \tag{43}$$

$$= 2^L \left(\frac{2e \cdot \mathfrak{m} \cdot \mathfrak{n} \cdot (\sum_{l=1}^{L} \mathscr{U}_l \cdot l)}{\sum_{l=1}^{L} Q_{(l)}}\right)^{\sum_{l=1}^{L} Q_{(l)}} \tag{44}$$

Next, we induce a transform to move the variable $\mathfrak{m}$ from the right side to the left side in (44), which is detailed in Lemma 18 of (Harvey et al., 2017). Suppose $2e \cdot \mathfrak{n} \cdot (\sum_{l=1}^{L} \mathscr{U}_l \cdot l) > 16$ and $\mathfrak{m} \geqslant \sum_{l=1}^{L} Q_{(l)} \geqslant L \geqslant 0$, which is easy to satisfy in practice because there are usually multiple nodes, units, and layers in an SGCN, and the sample size $\mathfrak{m}$ is much more larger than $\sum_{l=1}^{L} Q_{(l)}$ and $Q_{(l)} \geqslant 1$. Then (44) can be transformed into the following inequality by Lemma 18 of (Harvey et al., 2017):

$$\begin{aligned}\mathfrak{m} \leqslant & L + (\sum_{l=1}^{L} Q_{(l)}) \log_2(4e\mathfrak{n}(\sum_{l=1}^{L} \mathscr{U}_l l) \log_2(2e\mathfrak{n}(\sum_{l=1}^{L} \mathscr{U}_l l))) \\ = & \mathcal{O}(\bar{L} Q \log(\mathfrak{n}\mathscr{U})) \leqslant \mathcal{O}(LQ \log(\mathfrak{n}\mathscr{U})),\end{aligned} \tag{45}$$

where $\bar{L} := (\sum_{l=1}^{L} Q_{(l)})/Q \leqslant L$ is called the *effective depth* of a network. We let $\mathfrak{m} = \mathrm{Pdim}(\mathrm{SGCN})$ and observe that $\mathscr{U} \leqslant Q$, (19) has been proved.

□

A.1.5 Proof of Theorem 2

Proof. The core idea of the proof is to decompose the approximation of the empirical estimator $\hat{\theta}$ to the ground-truth parameter θ^* into two parts: (1) the approximation of $\hat{\theta}$ to the best approximation θ_T in the defined class of SGCNs and (2) the approximation of θ_T to θ^*:

$$\theta_T = \underset{\mathscr{G}_\theta \in \mathrm{SGCN}}{\mathrm{argmin}} [\|\theta_k - \theta_k^*\|_{L_\infty(X)}]_{k=1}^{d_\theta}, \quad \varepsilon_T = [\|\theta_{T,k} - \theta_k^*\|_{L_\infty(X)}]_{k=1}^{d_\theta} \in \mathbb{R}_+^{d_\theta}, \tag{46}$$

where argmin is implemented in an element-wise way w.r.t. the L_∞-norm on the probability space of X.

From Assumption 1 and that $\hat{\theta}$ optimizes (8),

$$\begin{aligned}
&c_1\mathbb{E}[\|\hat{\theta}(X)-\theta^*(X)\|_2^2]\\
\leqslant&\mathbb{E}[\ell(R,F,\mathscr{G}_{\hat{\theta}}(X))]-\mathbb{E}[\ell(R,F,\mathscr{G}_{\theta^*}(X))]\\
\leqslant&\underbrace{\mathbb{E}[\ell(R,F,\mathscr{G}_{\hat{\theta}}(X))]}_{①}-\underbrace{\mathbb{E}[\ell(R,F,\mathscr{G}_{\theta^*}(X))]}_{①}-\underbrace{\mathbb{E}_T[\ell(R,F,\mathscr{G}_{\hat{\theta}}(X))]}_{①}+\underbrace{\mathbb{E}_T[\ell(R,F,\mathscr{G}_{\theta_T}(X))]}_{②}\\
&+(\underbrace{\mathbb{E}_T[\ell(R,F,\mathscr{G}_{\theta^*}(X))]}_{①}-\underbrace{\mathbb{E}_T[\ell(R,F,\mathscr{G}_{\theta^*}(X))]}_{②})\\
=&\underbrace{(\mathbb{E}-\mathbb{E}_T)[\ell(R,F,\mathscr{G}_{\hat{\theta}}(X))-\ell(R,F,\mathscr{G}_{\theta^*}(X))]}_{①}+\underbrace{\mathbb{E}_T[\ell(R,F,\mathscr{G}_{\theta_T}(X))-\ell(R,F,\mathscr{G}_{\theta^*}(X))]}_{②}.
\end{aligned}\tag{47}$$

Part ②:

We first look into part ②, which corresponds to the bias term or approximation error. Let

$$\xi_t:=\frac{1}{T}\{(\ell(r_t,f_t,\mathscr{G}_{\theta_T}(x_t))-\ell(r_t,f_t,\mathscr{G}_{\theta^*}(x_t)))-\mathbb{E}[\ell(R,F,\mathscr{G}_{\theta_T}(X))-\ell(R,F,\mathscr{G}_{\theta^*}(X))]\},\quad t=1,\cdots,T.\tag{48}$$

Then $\{\xi_t\}_{t=1}^T$ are independent mean-zero random variables that satisfy the conditions of Lemma A2. Moreover,

$$\begin{aligned}
\nu_T^2=&\sum_{t=1}^T\mathbb{E}[\xi_t^2]\\
=&\frac{1}{T}\{\mathbb{E}[\ell(R,F,\mathscr{G}_{\theta_T}(X))-\ell(R,F,\mathscr{G}_{\theta^*}(X))]^2-(\mathbb{E}[\ell(R,F,\mathscr{G}_{\theta_T}(X))-\ell(R,F,\mathscr{G}_{\theta^*}(X))])^2\}\\
\leqslant&\frac{1}{T}\mathbb{E}[\ell(R,F,\mathscr{G}_{\theta_T}(X))-\ell(R,F,\mathscr{G}_{\theta^*}(X))]^2\\
\leqslant&\frac{1}{T}C_\ell^2\mathbb{E}[\|\theta_T(X)-\theta^*(X)\|_2^2]\\
\leqslant&\frac{1}{T}C_\ell^2\|\varepsilon_T\|_2^2.
\end{aligned}\tag{49}$$

On the other hand,

$$\begin{aligned}
|\xi_t|\leqslant&\frac{1}{T}\{|\ell(r_t,f_t,\mathscr{G}_{\theta_T}(x_t))-\ell(r_t,f_t,\mathscr{G}_{\theta^*}(x_t))|+\mathbb{E}[|\ell(R,F,\mathscr{G}_{\theta_T}(X))-\ell(R,F,\mathscr{G}_{\theta^*}(X))|]\}\\
\leqslant&\frac{C_\ell}{T}\{\|\theta_T(x_t)-\theta^*(x_t)\|_2+\mathbb{E}[\|\theta_T(X)-\theta^*(X)\|_2]\}\\
\leqslant&\frac{C_\ell}{T}\{\|\theta_T(x_t)\|_2+\|\theta^*(x_t)\|_2+\mathbb{E}[\|\theta_T(X)\|_2+\|\theta^*(X)\|_2]\}\\
\leqslant&\frac{C_\ell\sqrt{d_\theta}}{T}(2M+M+2M+M)\\
\leqslant&\frac{6C_\ell M\sqrt{d_\theta}}{T}.
\end{aligned}\tag{50}$$

Inserting (49) and (50) into Lemma A2, the following holds with probability at least $(1-\exp(-\zeta))$,

$$\begin{aligned}
&\mathbb{E}_T[\ell(R,F,\mathscr{G}_{\theta_T}(X))-\ell(R,F,\mathscr{G}_{\theta^*}(X))]\\
\leqslant&\mathbb{E}[\ell(R,F,\mathscr{G}_{\theta_T}(X))-\ell(R,F,\mathscr{G}_{\theta^*}(X))]+\frac{2\zeta\mathscr{M}}{3}+\sqrt{2\zeta}\nu_T\\
\leqslant&c_2\|\varepsilon_T\|_2^2+\frac{4\zeta C_\ell M\sqrt{d_\theta}}{T}+C_\ell\|\varepsilon_T\|_2\cdot\sqrt{\frac{2\zeta}{T}},
\end{aligned}\tag{51}$$

which bounds part ②.

Part ①:

In order to bound part ①, we need to exploit the theoretical results in Appendix A.1.2. Let

$$\mathscr{H} := \{g = \ell(r, f, \mathscr{G}_\theta(x)) - \ell(r, f, \mathscr{G}_{\theta^*}(x)) : \mathscr{G}_\theta \in \mathrm{SGCN}\}. \tag{52}$$

By following similar deductions to those of (49) and (50), we obtain

$$\mathbb{V}(\hat{g}) \leqslant \mathbb{E}[\hat{g}]^2 = \mathbb{E}[\ell(R, F, \mathscr{G}_{\hat{\theta}}(X)) - \ell(R, F, \mathscr{G}_{\theta^*}(X))]^2 \leqslant C_\ell^2 \iota^2 d_\theta, \tag{53}$$

$$|\hat{g}| = |\ell(r, f, \mathscr{G}_{\hat{\theta}}(x)) - \ell(r, f, \mathscr{G}_{\theta^*}(x))| \leqslant 3 C_\ell M \sqrt{d_\theta}, \tag{54}$$

where $\iota \geqslant \max_{1 \leqslant k \leqslant d_\theta} \|\hat{\theta}_k - \theta_k^*\|_{L_2(X)}$ is some initial upper bound for further tightening. Based on the boundedness assumptions w.r.t. θ^* and $\hat{\theta}$ from Assumption 1 and Theorem 2, this ι always exists, for example, $\iota = 3M\sqrt{d_\theta}$. Inserting (53) and (54) into (32) and using Lemma A5, the following holds with probability at least $(1 - 2\exp(-\zeta))$:

$$\begin{aligned}
&(\mathbb{E} - \mathbb{E}_T)[\ell(R, F, \mathscr{G}_{\hat{\theta}}(X)) - \ell(R, F, \mathscr{G}_{\theta^*}(X))] \\
\leqslant& 6 \underbrace{\mathbb{E}_\varsigma \mathcal{R}_T \mathscr{H}} + C_\ell \iota \cdot \sqrt{\frac{2\zeta d_\theta}{T}} + \frac{23\zeta C_\ell M \sqrt{d_\theta}}{T}.
\end{aligned} \tag{55}$$

The next step is to bound the empirical Rademacher complexity $\mathbb{E}_\varsigma \mathcal{R}_T \mathscr{H}$:

$$\begin{aligned}
&\mathbb{E}_\varsigma \mathcal{R}_T \mathscr{H} \\
=& \frac{1}{T} \mathbb{E}[\sup_{\mathscr{G}_\theta \in \mathrm{SGCN}} \sum_{t=1}^{T} \varsigma_t (\ell(r_t, f_t, \mathscr{G}_\theta(x_t)) - \ell(r_t, f_t, \mathscr{G}_{\theta^*}(x_t)))] && (56) \\
\leqslant& \frac{\sqrt{2} C_\ell}{T} \mathbb{E}[\sup_{\mathscr{G}_\theta \in \mathrm{SGCN}} \sum_{t=1}^{T} \sum_{k=1}^{d_\theta} \varsigma_{tk} (\theta_k(x_t) - \theta_k^*(x_t))] && (57) \\
\leqslant& \sqrt{2} C_\ell \sum_{k=1}^{d_\theta} \frac{1}{T} \mathbb{E}[\sup_{\mathscr{G}_\theta \in \mathrm{SGCN}} \sum_{t=1}^{T} \varsigma_{tk} (\theta_k(x_t) - \theta_k^*(x_t))] && (58) \\
=& \sqrt{2} C_\ell \sum_{k=1}^{d_\theta} \mathbb{E}_\varsigma \mathcal{R}_T \{\theta_k - \theta_k^* : \mathscr{G}_\theta \in \mathrm{SGCN}, \|\theta_k - \theta_k^*\|_{L_2(X)} \leqslant \iota\} && (59) \\
\leqslant& \sqrt{2} C_\ell \sum_{k=1}^{d_\theta} \mathbb{E}_\varsigma \mathcal{R}_T \{\theta_k - \theta_k^* : \mathscr{G}_\theta \in \mathrm{SGCN}, \|\theta_k - \theta_k^*\|_T \leqslant 2\iota\} && (60) \\
\leqslant& \sqrt{2} C_\ell d_\theta \inf_{0 \leqslant \varrho \leqslant 2\iota} \left\{ 4\varrho + \frac{12}{\sqrt{T}} \int_\varrho^{2\iota} \sqrt{\log \mathcal{E}(\rho, \mathrm{SGCN}, \|\cdot\|_T)}\, \mathrm{d}\rho \right\} && (61) \\
\leqslant& \sqrt{2} C_\ell d_\theta \inf_{0 \leqslant \varrho \leqslant 2\iota} \left\{ 4\varrho + \frac{12}{\sqrt{T}} \int_\varrho^{2\iota} \sqrt{\log \mathcal{E}(\rho, \mathrm{SGCN}|_{\{x_t\}_{t=1}^T}, L_\infty)}\, \mathrm{d}\rho \right\} && (62)
\end{aligned}$$

with probability at least $(1 - \exp(-\zeta))$.

Applying Lemma A4 to (62), we obtain

$$\begin{aligned}
&\inf_{0 \leqslant \varrho \leqslant 2\iota} \left\{ 4\varrho + \frac{12}{\sqrt{T}} \int_\varrho^{2\iota} \sqrt{\log \mathcal{E}(\rho, \mathrm{SGCN}|_{\{x_t\}_{t=1}^T}, L_\infty)}\, \mathrm{d}\rho \right\} \\
\leqslant& \inf_{0 \leqslant \varrho \leqslant 2\iota} \left\{ 4\varrho + \frac{12}{\sqrt{T}} \int_\varrho^{2\iota} \sqrt{\mathrm{Pdim}(\mathrm{SGCN}) \log \left(\frac{2eMT}{\rho \mathrm{Pdim}(\mathrm{SGCN})} \right)}\, \mathrm{d}\rho \right\} \\
\leqslant& 32\iota \sqrt{\frac{\mathrm{Pdim}(\mathrm{SGCN})}{T} \left(\log \frac{2eM}{\iota} + \frac{3}{2} \log T \right)}
\end{aligned} \tag{63}$$

when $T > \mathrm{Pdim}(\mathrm{SGCN})$ and $\varrho = 2\iota\sqrt{\frac{\mathrm{Pdim}(\mathrm{SGCN})}{T}} < 2\iota$. Hence, as long as $\iota \geqslant 1/T$ and $T \geqslant (2eM)^2$,

$$\mathbb{E}_\varsigma \mathcal{R}_T \mathscr{H} \leqslant 64\sqrt{2} C_\ell d_\theta \iota \sqrt{\frac{\mathrm{Pdim}(\mathrm{SGCN}) \cdot \log T}{T}}. \tag{64}$$

Inserting (64) into (55), we have

$$\begin{aligned}&(\mathbb{E}-\mathbb{E}_T)[\ell(R,F,\mathscr{G}_{\hat{\theta}}(X))-\ell(R,F,\mathscr{G}_{\theta^*}(X))]\\ \leqslant& C_{\ell,2}d_\theta\iota\sqrt{\frac{\mathrm{Pdim}(\mathrm{SGCN})\cdot\log T}{T}}+C_\ell\iota\cdot\sqrt{\frac{2\zeta d_\theta}{T}}+\frac{23\zeta C_\ell M\sqrt{d_\theta}}{T},\end{aligned}\tag{65}$$

where $C_{\ell,2}:=6\cdot 64\sqrt{2}C_\ell$.

Part ③:

In this part, we combine **Parts** ① and ② by inserting (51) and (65) into (47). Then the following holds with probability at least $(1-4\exp(-\zeta))$:

$$\begin{aligned}&c_1\mathbb{E}[\|\hat{\theta}(X)-\theta^*(X)\|_2^2]\\ \leqslant& C_{\ell,2}d_\theta\iota\sqrt{\frac{\mathrm{Pdim}(\mathrm{SGCN})\cdot\log T}{T}}+C_\ell\iota\cdot\sqrt{\frac{2\zeta d_\theta}{T}}+\frac{23\zeta C_\ell M\sqrt{d_\theta}}{T}\\ &+c_2\|\varepsilon_T\|_2^2+\frac{4\zeta C_\ell M\sqrt{d_\theta}}{T}+C_\ell\|\varepsilon_T\|_2\cdot\sqrt{\frac{2\zeta}{T}}\qquad(66)\\ \leqslant&\left(C_{\ell,2}d_\theta\sqrt{\frac{C_dQL\log(Q\mathfrak{n})\cdot\log T}{T}}\right)\cdot\iota+\left(C_\ell\sqrt{\frac{2\zeta d_\theta}{T}}\right)\cdot\iota+\frac{27\zeta C_\ell M\sqrt{d_\theta}}{T}\\ &+c_2\|\varepsilon_T\|_2^2+C_\ell\|\varepsilon_T\|_2\cdot\sqrt{\frac{2\zeta}{T}}.\qquad(67)\end{aligned}$$

The last inequality holds by Theorem 1. As long as $\|\varepsilon_T\|_2\ll\iota$ and $\sqrt{\frac{QL\log(Q\mathfrak{n})\cdot\log T}{T}}\ll\iota$, then the prior knowledge that $\iota_0^2\leqslant\mathbb{E}[\|\hat{\theta}(X)-\theta^*(X)\|_2^2]\leqslant\mathfrak{t}_0^2$ implies $\iota_1^2\leqslant\mathbb{E}[\|\hat{\theta}(X)-\theta^*(X)\|_2^2]\leqslant\mathfrak{t}_1^2$ with high probability for $\iota_1\leqslant\mathfrak{t}_1\ll\mathfrak{t}_0/\sqrt{d_\theta}\leqslant\iota_0$. In this way, we can progressively tighten the bound ι_0 to a fixed point or critical radius ι_*.

Specifically, define ι_* as follows:

$$\iota_*=\inf\{\iota>0:18M\mathbb{E}\mathcal{R}_T\{\theta_k-\theta_k^*:\mathscr{G}_\theta\in\mathrm{SGCN},\|\theta_k-\theta_k^*\|_{L_2(X)}\leqslant\mathfrak{s},\forall k\}<\mathfrak{s}^2,\forall\mathfrak{s}\geqslant\iota\}.\tag{68}$$

It can be seen that both ι_* and $2\iota_*$ satisfy (38). Define the following data-dependent event:

$$E:=\{\|\theta_k-\theta_k^*\|_T\leqslant 4\iota_*,\forall\mathscr{G}_\theta\in\mathrm{SGCN}\text{ and }\|\theta_k-\theta_k^*\|_{L_2(X)}\leqslant 2\iota_*,\forall k\}\tag{69}$$

and denote $\mathbb{1}_E$ as its indicator function. Then we have $\mathbb{P}\{\mathbb{1}_E=1\}\geqslant(1-1/T)$ if $\iota_*\geqslant\sqrt{18}M\sqrt{\log T/T}$, based on (39) and (40).

Next, we can bound ι_* via Theorem 1 and Lemma A5:

$$\begin{aligned}\iota_*^2\leqslant&18M\mathbb{E}\mathcal{R}_T\{\theta_k-\theta_k^*:\mathscr{G}_\theta\in\mathrm{SGCN},\|\theta_k-\theta_k^*\|_{L_2(X)}\leqslant\iota_*,\forall k\}\\ \leqslant&18M\mathbb{E}\mathcal{R}_T\{\theta_k-\theta_k^*:\mathscr{G}_\theta\in\mathrm{SGCN},\|\theta_k-\theta_k^*\|_{L_2(X)}\leqslant 2\iota_*,\forall k\}\\ \leqslant&18M\mathbb{E}\{\mathbb{E}_\varsigma\mathcal{R}_T\{\theta_k-\theta_k^*:\mathscr{G}_\theta\in\mathrm{SGCN},\|\theta_k-\theta_k^*\|_T\leqslant 4\iota_*,\forall k\}\cdot\mathbb{1}_E+3M(1-\mathbb{1}_E)\}\\ \leqslant&36MC_{\ell,2}\iota_*\sqrt{\frac{C_dQL\log(Q\mathfrak{n})\cdot\log T}{T}}+54M^2\frac{1}{T}\\ \leqslant&72MC_{\ell,2}\iota_*\sqrt{\frac{C_dQL\log(Q\mathfrak{n})\cdot\log T}{T}},\end{aligned}\tag{70}$$

where the last inequality holds with $\iota_*\geqslant\sqrt{18}M\sqrt{\log T/T}$ and a sufficiently large T. Dividing both sides of (70) by ι_*, we have

$$\iota_*\leqslant 72MC_{\ell,2}\sqrt{\frac{C_dQL\log(Q\mathfrak{n})\cdot\log T}{T}}.\tag{71}$$

Part ④:

In this part, we dig into the core of localization by alternating Lemma A7 and **Part ③**. We can divide the function space SGCN into shells by iteratively excluding L_2 balls with increasing radii:

$$\mathcal{B}(\theta_k^*, \bar{\iota}), \mathcal{B}(\theta_k^*, 2\bar{\iota})\backslash\mathcal{B}(\theta_k^*, \bar{\iota}), \cdots, \mathcal{B}(\theta_k^*, 2^{\mathfrak{l}}\bar{\iota})\backslash\mathcal{B}(\theta_k^*, 2^{\mathfrak{l}-1}\bar{\iota}), \tag{72}$$

where $\mathfrak{l}$ is chosen as the largest positive integer no larger than $\log_2(\frac{2M}{\sqrt{(\log T)/T}})$. Localization aims to find a suitable bound on $\bar{\iota}$.

First, suppose $\bar{\iota} > \iota_*$. Then for each shell j, Lemma A7 indicates that with probability at least $(1 - 2\mathfrak{l}\exp(-\zeta))$,

$$\|\theta_k - \theta_k^*\|_{L_2(X)} \leqslant 2^{\mathrm{j}}\bar{\iota} \quad \Rightarrow \quad \|\theta_k - \theta_k^*\|_T \leqslant 2^{\mathrm{j}+1}\bar{\iota}, \quad \forall k. \tag{73}$$

If for some $\mathrm{j} \leqslant \mathfrak{l}$ and k,

$$\hat{\theta}_k \in \mathcal{B}(\theta_k^*, 2^{\mathrm{j}}\bar{\iota})\backslash\mathcal{B}(\theta_k^*, 2^{\mathrm{j}-1}\bar{\iota}), \tag{74}$$

then we can apply (67) to this specific dimension k and set $d_\theta = 1$. If we choose

$$\begin{aligned}\bar{\iota} =& \frac{8}{c_1}\left(C_{\ell,2}\sqrt{\frac{C_d Q L \log(Q\mathfrak{n})\cdot\log T}{T}} + C_\ell\sqrt{\frac{2\zeta}{T}}\right) \\ &+ \left(\sqrt{\frac{2(c_2 \vee 1)}{c_1}}\varepsilon_{T,k} + \sqrt{\frac{120\zeta C_\ell M}{c_1 T}}\right) + \iota_*,\end{aligned} \tag{75}$$

then the following two conditions are satisfied:

$$\begin{aligned}\frac{1}{c_1}\left(C_{\ell,2}\sqrt{\frac{C_d Q L \log(Q\mathfrak{n})\cdot\log T}{T}} + C_\ell\sqrt{\frac{2\zeta}{T}}\right) \leqslant& \frac{1}{8}2^{\mathrm{j}}\bar{\iota}, \\ \frac{1}{c_1}\left(\frac{27\zeta C_\ell M}{T} + c_2\varepsilon_{T,k}^2 + C_\ell\varepsilon_{T,k}\cdot\sqrt{\frac{2\zeta}{T}}\right) \leqslant& \frac{1}{8}2^{2\mathrm{j}}\bar{\iota}^2.\end{aligned} \tag{76}$$

Moreover,

$$\begin{aligned}&\|\hat{\theta}_k - \theta_k^*\|_{L_2(X)}^2 \\ \leqslant& \frac{1}{c_1}\left\{\left(C_{\ell,2}\sqrt{\frac{C_d Q L \log(Q\mathfrak{n})\cdot\log T}{T}} + C_\ell\sqrt{\frac{2\zeta}{T}}\right)\cdot 2^{\mathrm{j}}\bar{\iota} + \frac{27\zeta C_\ell M}{T} + c_2\varepsilon_{T,k}^2 + C_\ell\varepsilon_{T,k}\cdot\sqrt{\frac{2\zeta}{T}}\right\} \\ \leqslant& 2^{2\mathrm{j}-2}\bar{\iota}^2.\end{aligned} \tag{77}$$

Repeating the above tightening step from the $\mathfrak{l}$-th to the zeroth shells, we have that

$$\|\hat{\theta}_k - \theta_k^*\|_{L_2(X)} \leqslant 2^{\mathfrak{l}}\bar{\iota} \text{ and } \|\hat{\theta}_k - \theta_k^*\|_T \leqslant 2^{\mathfrak{l}+1}\bar{\iota} \tag{78}$$

implies

$$\|\hat{\theta}_k - \theta_k^*\|_{L_2(X)} \leqslant 2^0\bar{\iota} \text{ and } \|\hat{\theta}_k - \theta_k^*\|_T \leqslant 2^1\bar{\iota} \tag{79}$$

with probability at least $(1 - 6\mathfrak{l}\exp(-\zeta))$. We let $\tilde{\zeta} = \zeta - \log(6\mathfrak{l})$, recall $\mathfrak{l} \leqslant \log_2(\frac{2M}{\sqrt{(\log T)/T}})$, and combine (75) and (71). This leads to

$$\begin{aligned}\bar{\iota} \leqslant& \frac{8}{c_1}\left(C_{\ell,2}\sqrt{\frac{C_d Q L \log(Q\mathfrak{n})\cdot\log T}{T}} + C_\ell\sqrt{\frac{2(\log\log T + \tilde{\zeta})}{T}}\right) \\ &+ \left(\sqrt{\frac{2(c_2 \vee 1)}{c_1}}\varepsilon_{T,k} + \sqrt{\frac{120 C_\ell M(\log\log T + \tilde{\zeta})}{c_1 T}}\right) + \iota_* \\ \leqslant& C_{\ell,3}\left(\sqrt{\frac{Q L \log(Q\mathfrak{n})\cdot\log T}{T}} + \sqrt{\frac{\log\log T + \tilde{\zeta}}{T}} + \varepsilon_{T,k}\right)\end{aligned} \tag{80}$$

for some constant $C_{\ell,3} > 0$ that is independent of T. Inserting (80) into (79), we have actually proven the nonasymptotic bounds of estimators $\{\hat{\theta}_k\}_{k=1}^{d_\theta}$ for SGCN with Q, L, $\mathfrak{n}$, T, and $\tilde{\zeta}$ (holds with probability at least $(1-\exp(-\tilde{\zeta}))$).

Part ⑤:

In this final part, we specify suitable choices for J, L, and $\varepsilon_{T,k}$. For the function class of SGCN, we have $Q \leqslant \mathcal{C}_0 J^2 L$ with some constant $\mathcal{C}_0$. Inserting this into (80) yields

$$\bar{\iota} \leqslant \mathcal{C}_1 \left(\sqrt{\frac{J^2 L^2 \log(J^2 L \mathfrak{n}) \cdot \log T}{T}} + \sqrt{\frac{\log\log T + \tilde{\zeta}}{T}} + \varepsilon_{T,k} \right). \tag{81}$$

Next, we invoke Lemma A6 with $\mathscr{F}_{DNN} := \mathrm{SGCN}$, $q = d_c$, and $\mathrm{Pdim}(\mathrm{SGCN})$ in Theorem 1 (already inserted into Eq. 81) to obtain the scalable approximation property of SGCN:

$$\begin{aligned} J(\varepsilon_{T,k}) \leqslant & Q(\varepsilon_{T,k}) L(\varepsilon_{T,k}) \leqslant \mathcal{C}_0^2 \varepsilon_{T,k}^{-\frac{d_c}{p}} (\log(\frac{1}{\varepsilon_{T,k}}) + 1)^2, \\ L(\varepsilon_{T,k}) \leqslant & \mathcal{C}_0 (\log(\frac{1}{\varepsilon_{T,k}}) + 1). \end{aligned} \tag{82}$$

By absorbing the terms of lower orders, we have

$$J^2 L^2 \log(J^2 L \mathfrak{n}) \leqslant \tilde{\mathcal{C}} \varepsilon_{T,k}^{-\frac{2d_c}{p}} (\log(\frac{1}{\varepsilon_{T,k}}) + 1)^6 (\log(\frac{\mathfrak{n}}{\varepsilon_{T,k}}) + 1). \tag{83}$$

Hence we can choose $\varepsilon_{T,k} = T^{-\frac{p}{2(p+d_c)}}$, $J \asymp T^{d_c/2(p+d_c)} \log^2 T$, and $L \asymp \log T$. Then (81) is transformed into

$$\bar{\iota} \leqslant \mathcal{C}_3 \left(T^{-\frac{p}{2(p+d_c)}} \log^{\frac{7}{2}} T \cdot \log^{\frac{1}{2}}(\mathfrak{n} T) + \sqrt{\frac{\log\log T + \tilde{\zeta}}{T}} \right). \tag{84}$$

Last, we combine (79) and (84), and absorb the terms of lower orders. Then the following holds with probability at least $(1-\exp(-\tilde{\zeta}))$:

$$\|\hat{\theta}_k - \theta_k^*\|_{L_2(X)}^2 \leqslant \bar{\iota}^2 \leqslant \mathcal{C} \left(T^{-\frac{p}{p+d_c}} \log^7 T \cdot \log(\mathfrak{n} T) + \frac{\log\log T + \tilde{\zeta}}{T} \right), \tag{85}$$

$$\|\hat{\theta}_k - \theta_k^*\|_T^2 \leqslant 4\bar{\iota}^2 \leqslant 4\mathcal{C} \left(T^{-\frac{p}{p+d_c}} \log^7 T \cdot \log(\mathfrak{n} T) + \frac{\log\log T + \tilde{\zeta}}{T} \right). \tag{86}$$

This finishes the entire proof of Theorem 2.

□

A.2 Proof of Theorem 3

The key to establish the asymptotic normality is to construct the influence function via a standard Neyman orthogonalization approach. We index distributions and other nonparametric objects by η, and define $\theta^*(x;\eta)$ and $\mu^*(\eta)$ as follows:

$$\theta^*(x;\eta) := \operatorname*{argmin}_{\mathfrak{c}} \int \ell(\mathfrak{w}, \mathfrak{c}(x)) \mathfrak{h}_{\mathfrak{w}}(\mathfrak{w};\eta)\, \mathrm{d}\mathfrak{w}, \tag{87}$$

$$\mu^*(\eta) := \int \mathcal{F}(x, \theta(x;\eta); \tilde{f}) \mathfrak{h}_x(x;\eta)\, \mathrm{d}x, \tag{88}$$

where $\mathfrak{h}_{\mathfrak{w}}$ and $\mathfrak{h}_x$ denote the density functions of $\mathfrak{w} := (r, f, x)$ and x (assumed to exist), respectively. To simplify notation, we directly denote $\theta(x;\eta)$ as $\mathscr{G}_\theta(x;\eta)$ inside $\mathcal{F}$. $\eta = 0$ represents the true data generating process, where η itself can be omitted: $\mathfrak{h}_{\mathfrak{w}}(\mathfrak{w};0) = \mathfrak{h}_{\mathfrak{w}}(\mathfrak{w})$ and $\theta^*(x;0) = \theta^*(x)$.

The objective is to find a tractable pathwise derivative decomposition as follows:

$$\frac{\partial \mu(\eta)}{\partial \eta}|_{\eta=0} := \mathbb{E}[\psi(\mathfrak{W})\mathscr{S}(\mathfrak{W})], \quad \mathscr{S}(\mathfrak{w}) = \mathscr{S}(\mathfrak{w};\eta)|_{\eta=0}. \tag{89}$$

To do this, we differentiate (88) w.r.t. η and let $\eta = 0$, which leads to:

$$\begin{aligned}
&\frac{\partial \mu(\eta)}{\partial \eta}|_{\eta=0} \\
:=&\frac{\partial}{\partial \eta}\left\{\int \mathcal{F}(x,\theta(x;\eta);\tilde{f})\mathfrak{h}_x(x;\eta)\,\mathrm{d}x\right\}|_{\eta=0} \\
=&\int \mathcal{F}(x,\theta(x;0);\tilde{f})\frac{\partial \mathfrak{h}_x(x;\eta)}{\partial \eta}|_{\eta=0}\,\mathrm{d}x + \int \frac{\partial \mathcal{F}(x,\theta(x;\eta);\tilde{f})}{\partial \eta}|_{\eta=0}\mathfrak{h}_x(x;0)\,\mathrm{d}x \\
=&\underbrace{\int \mathcal{F}(x,\theta^*(x);\tilde{f})\frac{\partial \mathfrak{h}_x(x;\eta)}{\partial \eta}|_{\eta=0}\,\mathrm{d}x}_{\text{Distribution Change}} + \underbrace{\int \mathcal{F}_{[\theta]}(x,\theta^*(x);\tilde{f})\theta_{[\eta]}(x)\mathfrak{h}_x(x)\,\mathrm{d}x}_{\text{Parameter Change}}.
\end{aligned} \tag{90}$$

$\mathcal{F}_{[\theta]} \in \mathbb{R}^{d_\mu \times d_\theta}$ and $\theta_{[\eta]} \in \mathbb{R}^{d_\theta}$ are partial derivatives:

$$(\mathcal{F}_{[\theta]}(x,\theta^*(x);\tilde{f}))_{i,j} = \frac{\partial \mathcal{F}_i(x,\mathfrak{c};\tilde{f})}{\partial \mathfrak{c}_j}|_{\mathfrak{c}=\theta(x;0)}, \quad \theta_{[\eta]}(x) := \theta_{[\eta]}(x;0) = \frac{\partial \theta(x;\eta)}{\partial \eta}|_{\eta=0}, \tag{91}$$

where $(\mathcal{F}_{[\theta]})_{i,j}$ denotes the i,j-th element of the Jacobian $\mathcal{F}_{[\theta]}$.

Next, we need to reformulate both integrals in (90) as expectations w.r.t. the full score function $\mathscr{S}(\mathfrak{w})$. An eligible score function has the mean zero property and satisfies the logarithmic conditional probability:

$$\mathbb{E}[\mathscr{S}(R,F|X)|X] = 0, \tag{92}$$

$$\mathscr{S}(r,x,f) = \mathscr{S}(r,f|x) + \mathscr{S}(x). \tag{93}$$

Then the first integral in (90) can be transformed as

$$\begin{aligned}
&\underbrace{\int \mathcal{F}(x,\theta^*(x);\tilde{f})\frac{\partial \mathfrak{h}_x(x;\eta)}{\partial \eta}|_{\eta=0}\,\mathrm{d}x}_{\text{Distribution Change}} \\
=&\int \mathcal{F}(x,\theta^*(x);\tilde{f})\mathscr{S}(x)\mathfrak{h}_x(x)\,\mathrm{d}x \\
=&\mathbb{E}[\mathcal{F}(X,\theta^*(x);\tilde{f})\mathscr{S}(X)] \\
=&\mathbb{E}[\mathcal{F}(X,\theta^*(x);\tilde{f})\cdot(\mathscr{S}(R,X,F) - \mathscr{S}(R,F|X))] \\
=&\mathbb{E}[\mathcal{F}(X,\theta^*(x);\tilde{f})\mathscr{S}(R,X,F)] - \mathbb{E}[\mathcal{F}(X,\theta^*(x);\tilde{f})\cdot\mathbb{E}[\mathscr{S}(R,F|X)|X]] \\
=&\mathbb{E}[\mathcal{F}(X,\theta^*(x);\tilde{f})\mathscr{S}(R,X,F)].
\end{aligned} \tag{94}$$

It interprets the plug-in component of the influence function, which captures the sensitivity of $\mu(\eta)$ at $\eta = 0$ as if $\theta^*(x)$ were known.

On the other hand, the second integral in (90) serves as a correction term that accounts for the nonparametric estimation of the nuisance function. To do this, a crucial step is to find $\theta_{[\eta]}(x) = \frac{\partial \theta(x;\eta)}{\partial \eta}|_{\eta=0}$, which can be done via the following first-order condition:

$$\mathbb{E}_\eta[\ell_{[\theta]}(\mathfrak{W},\theta(x;\eta))|X = x] \equiv 0. \tag{95}$$

Without loss of generality, this condition can be interpreted as that the **maximum likelihood estimation** of θ should be **obtained** at $\ell_{[\theta]} = 0$. Specifically,

$$\ell_{[\theta]}(\mathfrak{w},\theta(x;\eta)) = \frac{\partial \ell(\mathfrak{w},\mathfrak{c})}{\partial \mathfrak{c}}|_{\mathfrak{c}=\theta(x;\eta)}. \tag{96}$$

Therefore, (95) is essentially:

$$\int \frac{\partial \ell(\mathfrak{w}, \mathfrak{c})}{\partial \mathfrak{c}}|_{\mathfrak{c}=\theta(x;\eta)} \mathfrak{h}_{r,f|x}(r, f; \eta|x)\, \mathrm{d}r\, \mathrm{d}f \equiv 0. \tag{97}$$

Differentiating both sides of (97) w.r.t. η yields:

$$\begin{aligned} &\int \frac{\partial \ell(\mathfrak{w}, \mathfrak{c})}{\partial \mathfrak{c}}|_{\mathfrak{c}=\theta(x;\eta)} \frac{\partial \mathfrak{h}_{r,f|x}(r, f; \eta|x)}{\partial \eta}\, \mathrm{d}r\, \mathrm{d}f \\ &+ \int \ell_{[\theta\theta]}(\mathfrak{w}, \theta(x;\eta))\theta_{[\eta]}(x;\eta)\mathfrak{h}_{r,f|x}(r, f; \eta|x)\, \mathrm{d}r\, \mathrm{d}f = 0, \end{aligned} \tag{98}$$

where $\ell_{[\theta\theta]}(\mathfrak{w}, \theta(x;\eta))$ denotes the Hessian matrix of ℓ w.r.t. θ:

$$(\ell_{[\theta\theta]}(\mathfrak{w}, \theta(x;\eta)))_{i,j} = \frac{\partial^2 \ell(\mathfrak{w}, \mathfrak{c})}{\partial \mathfrak{c}_i \partial \mathfrak{c}_j}|_{\mathfrak{c}=\theta(x;\eta)}. \tag{99}$$

Let $\eta = 0$ in (98), we have

$$\begin{aligned} &\int \ell_{[\theta]}(\mathfrak{w}, \theta^*(x)) \left[\frac{\partial \mathfrak{h}_{r,f|x}(r, f; \eta|x)}{\partial \eta}|_{\eta=0}\right] \mathrm{d}r\, \mathrm{d}f \\ &+ \int \ell_{[\theta\theta]}(\mathfrak{w}, \theta^*(x))\theta_{[\eta]}(x)\mathfrak{h}_{r,f|x}(r, f; 0|x)\, \mathrm{d}r\, \mathrm{d}f \end{aligned} \tag{100}$$

$$\begin{aligned} =&\int \ell_{[\theta]}(\mathfrak{w}, \theta^*(x)) \left[\mathscr{S}(r, f|x)\mathfrak{h}_{r,f|x}(r, f; 0|x)\right] \mathrm{d}r\, \mathrm{d}f \\ &+ \int \ell_{[\theta\theta]}(\mathfrak{w}, \theta^*(x))\theta_{[\eta]}(x)\mathfrak{h}_{r,f|x}(r, f; 0|x)\, \mathrm{d}r\, \mathrm{d}f = 0, \end{aligned} \tag{101}$$

where the equality inside the brackets holds due to the definition and the chain rule of the score function $\mathscr{S}$. (101) can be represented as

$$\mathbb{E}[\ell_{[\theta]}(\mathfrak{W}, \theta^*(x))\mathscr{S}(R, F|X)|X] + \mathbb{E}[\ell_{[\theta\theta]}(\mathfrak{W}, \theta^*(x))\theta_{[\eta]}(x)|X] = 0. \tag{102}$$

Since θ only depends on X, (102) can be transformed as

$$\mathbb{E}[\ell_{[\theta\theta]}(\mathfrak{W}, \theta^*(x))|X]\theta_{[\eta]}(x) = -\mathbb{E}[\ell_{[\theta]}(\mathfrak{W}, \theta^*(x))\mathscr{S}(R, F|X)|X]. \tag{103}$$

From Assumption 3(3), $\Lambda(x) := \mathbb{E}[\ell_{[\theta\theta]}(\mathfrak{W}, \theta^*(x))|X = x]$ is invertible, thus we can use (103) to solve for $\theta_{[\eta]}(x)$:

$$\theta_{[\eta]}(x) = -\mathbb{E}[\Lambda(X)^{-1}\ell_{[\theta]}(\mathfrak{W}, \theta^*(x))\mathscr{S}(R, F|X)|X]. \tag{104}$$

Substituting (104) into the second integral of (90) leads to

$$\begin{aligned} &\underbrace{\int \mathcal{F}_{[\theta]}(x, \theta^*(x); \tilde{f})\theta_{[\eta]}(x)\mathfrak{h}_x(x)\, \mathrm{d}x}_{\text{Parameter Change}} \\ =& -\mathbb{E}[\mathcal{F}_{[\theta]}(X, \theta^*(X); \tilde{f}) \cdot \mathbb{E}[\Lambda(X)^{-1}\ell_{[\theta]}(\mathfrak{W}, \theta^*(x))\mathscr{S}(R, F|X)|X]] \\ =& -\mathbb{E}[\mathbb{E}[\mathcal{F}_{[\theta]}(X, \theta^*(X); \tilde{f}) \cdot \Lambda(X)^{-1}\ell_{[\theta]}(\mathfrak{W}, \theta^*(x))\mathscr{S}(R, F|X)|X]] \\ =& -\mathbb{E}[\mathcal{F}_{[\theta]}(X, \theta^*(X); \tilde{f}) \cdot \Lambda(X)^{-1}\ell_{[\theta]}(\mathfrak{W}, \theta^*(x))\mathscr{S}(R, F|X)]. \end{aligned} \tag{105}$$

Using (93) and (95) in (105) and applying a similar transform to (94), we have

$$\begin{aligned} &-\mathbb{E}[\mathcal{F}_{[\theta]}(X, \theta^*(X); \tilde{f}) \cdot \Lambda(X)^{-1}\ell_{[\theta]}(\mathfrak{W}, \theta^*(x))[\mathscr{S}(R, X, F) - \mathscr{S}(X)]] \\ =&\mathbb{E}[\mathcal{F}_{[\theta]}(X, \theta^*(X); \tilde{f}) \cdot \Lambda(X)^{-1}\ell_{[\theta]}(\mathfrak{W}, \theta^*(x))\mathscr{S}(X)] \\ &-\mathbb{E}[\mathcal{F}_{[\theta]}(X, \theta^*(X); \tilde{f}) \cdot \Lambda(X)^{-1}\ell_{[\theta]}(\mathfrak{W}, \theta^*(x))\mathscr{S}(R, X, F)] \\ =&\mathbb{E}[\mathcal{F}_{[\theta]}(X, \theta^*(X); \tilde{f}) \cdot \Lambda(X)^{-1}\mathbb{E}[\ell_{[\theta]}(\mathfrak{W}, \theta^*(x))|X]\mathscr{S}(X)] \\ &-\mathbb{E}[\mathcal{F}_{[\theta]}(X, \theta^*(X); \tilde{f}) \cdot \Lambda(X)^{-1}\ell_{[\theta]}(\mathfrak{W}, \theta^*(x))\mathscr{S}(R, X, F)] \end{aligned}$$

$$= - \mathbb{E}[\mathcal{F}_{[\theta]}(X, \theta^*(X); \tilde{f}) \cdot \Lambda(X)^{-1} \ell_{[\theta]}(\mathfrak{W}, \theta^*(x)) \mathscr{S}(R, X, F)]. \tag{106}$$

Substituting (94) and (106) into (90), we have

$$\begin{aligned} &\frac{\partial \mu(\eta)}{\partial \eta}|_{\eta=0} \\ =&\mathbb{E}[\mathcal{F}(X, \theta^*(x); \tilde{f}) \mathscr{S}(R, X, F)] - \mathbb{E}[\mathcal{F}_{[\theta]}(X, \theta^*(X); \tilde{f}) \cdot \Lambda(X)^{-1} \ell_{[\theta]}(\mathfrak{W}, \theta^*(x)) \mathscr{S}(R, X, F)]. \end{aligned} \tag{107}$$

Hence we can directly define $\psi(\mathfrak{w})$ as follows to satisfy (89):

$$\psi(\mathfrak{w}) := \mathcal{F}(x, \theta^*(x); \tilde{f}) - \mathcal{F}_{[\theta]}(x, \theta^*(x); \tilde{f}) \cdot \Lambda(x)^{-1} \ell_{[\theta]}(\mathfrak{w}, \theta^*(x)), \tag{108}$$

which is exactly the debiased version of μ defined in (13). Moreover, $(\psi - \mu^*)$ is the desired influence function that still satisfies (89), because $\mathbb{E}[\mu^* \mathscr{S}(\mathfrak{W})] = \mu^* \mathbb{E}[\mathscr{S}(\mathfrak{W})] = 0$.

The last step is to employ the standard results of DML (Chernozhukov et al., 2018). Assumption 3 in this paper ensures that the influence function $(\psi - \mu^*)$ satisfies the corresponding Assumptions 3.1 and 3.2 in (Chernozhukov et al., 2018): such as the (conditional) mean zero property, linearity, smoothness, Neyman orthogonality, eligible whitening transform $\hat{\Psi}^{-\frac{1}{2}}$ with $\hat{\Psi}$ defined in (15), moment conditions, and the nonasymptotic bound of parameter estimation derived in Theorem 2. Therefore, Theorem 3 in this paper also holds from the corresponding Theorems 3.1 and 3.2 in (Chernozhukov et al., 2018).

A.3 Experimental Setup

A.3.1 Data Set

Our empirical analysis is based on an extensive panel of U.S. equities spanning from 1957 to 2021, which consists of ordinary common shares listed on the NYSE, AMEX, and NASDAQ. The number of assets $\mathfrak{n}$ varies from about 1000 to 10000 during different periods because the listed assets are changing from 1957 to 2021, and this $\mathfrak{n}$ equals the number of nodes for the proposed SGCN. The conditioning information set at month t integrates both micro-level firm characteristics (heterogeneous instruments $Z_{i,t}$) and macro-level economic state variables (public instruments Y_t). Specifically, each stock i is characterized by a 95-dimensional vector of firm-specific attributes $Z_{i,t}$, supplemented by industry designations derived from the Fama-French 49-industry (FF49) classification (Fama & French, 1997). The primary predictive target is the one-month-ahead individual stock excess return $r_{i,t+1} := (R_{i,t+1} - r^f_{t+1})$, where $R_{i,t+1}$ and r^f_{t+1} denote the raw stock-i return and the risk-free return, respectively. To eliminate look-ahead bias, $Z_{i,t}$ is constructed only on the information observable by the end of month t. Furthermore, we implement a monthly cross-sectional rank normalization to map each feature into the $[-1, 1]$ interval, which ensures complete isolation from future cross-sectional distributions.

To capture time-varying macroeconomic conditions, we incorporate 10 aggregate predictors from (Welch & Goyal, 2008). We ensure temporal integrity by aligning these variables with their public release schedules: market-based indicators (e.g., term and default yield spreads) are lagged by one month, while indicators subject to official reporting delays (e.g., inflation) are lagged by two months. Finally, to ensure strict out-of-sample compliance, these macroeconomic predictors are standardized via a 15-year backward-looking rolling window, which ensures that normalization parameters depend exclusively on historical data. A concise summary of the data set construction and variable processing is provided in Table A1.

A.3.2 Details on Evaluating Metrics

The Predictive R^2 (Gu et al., 2020) is a widely-used metric that evaluates the generalizing capability of a model in the regression task, which extends the traditional R^2 to the out-of-sample scenario. Two types of Predictive R^2 are considered: the stock-level R^2_r and the portfolio-level R^2_{x}. R^2_r directly uses the ordinary return r_{t+1} and the predicted return $\hat{r}^{pred}_{t+1} = \hat{\Gamma}_\alpha(\hat{G}_\beta(z_t), \hat{G}_\gamma(y_t)) + \hat{G}_\beta(z_t)\hat{f}_t$, where $\hat{f}_t$ is the out-of-sample expected factor premium, typically estimated by the historical sample average of the latent factors $\hat{f}$ over a rolling window (e.g., from $(t - 60)$ to t). On the other hand,

Table A1: Summary statistics of the data set used in experiments.

Statistic	**Value**
Sample period	1957–2021
Time frequency	Monthly
Asset universe	U.S. common stocks (NYSE/AMEX/NASDAQ)
Number of assets	About 1000 $\sim$ 10000
Prediction horizon	One month ahead
Target variable	$r_{i,t+1} := (R_{i,t+1} - r^f_{t+1})$
Micro feature count	94
Marco feature count	10
Feature normalization (micro)	Cross-sectional rank normalization

R^2_{x} uses the characteristic-managed portfolios $\mathrm{x}_{t+1} = (z_t^\top z_t)^{-1} z_t^\top r_{t+1}$ and its predicted version $\hat{\mathrm{x}}^{pred}_{t+1} = (z_t^\top z_t)^{-1} z_t^\top \hat{r}^{pred}_{t+1}$. This isolates the predictive power directly associated with the anomaly characteristics z_t.

$$R^2_r = 1 - \frac{\sum_{i,t}(r_{i,t+1} - \hat{r}^{pred}_{i,t+1})^2}{\sum_{i,t} r^2_{i,t+1}}, \quad R^2_{\mathrm{x}} = 1 - \frac{\sum_{i,t}(\mathrm{x}_{i,t+1} - \hat{\mathrm{x}}^{pred}_{i,t+1})^2}{\sum_{i,t} \mathrm{x}^2_{i,t+1}}. \tag{109}$$

Since the historical mean $\bar{r}$ (or $\bar{\mathrm{x}}$) is usually very close to 0 in practice, it can be directly set to $\bar{r} = 0$ (or $\bar{\mathrm{x}} = 0$) inside the square of the denominator in the asset pricing literature.

Besides the pure return, the risk of an investing strategy is also very important in asset pricing, since investors require compensation for uncertainty in a rational market. The SR is a mainstream risk-adjusted return that measures such an investing efficiency rather than the absolute performance. It can be defined as follows:

$$SR = \frac{\bar{r}_{str} - \bar{r}^f}{\hat{\sigma}\left(r_{str}\right)}, \tag{110}$$

where $\bar{r}_{str}$ denotes the out-of-sample historical mean return of some investing strategy, $\hat{\sigma}\left(r_{str}\right)$ denotes the corresponding standard deviation (STD), and $\bar{r}^f$ denotes the historical mean risk-free return.

A.4 Implementation Details

All the neural network architectures and experiments are implemented via PyTorch[1] with an NVIDIA RTX A6000 GPU. Model parameters are optimized using the Adam optimizer with a conservative initial learning rate of 1×10^{-4} and an L_2 regularization strength (i.e., weight decay) of 1×10^{-4}. To enhance robustness against market noise, we apply a dropout rate of 0.5 across all hidden layers and use gradient clipping with a maximum norm threshold of 1.0. The optimization objective is the Mean Squared Error (MSE) loss with training proceeding for a maximum of 200 epochs. An early stopping mechanism is enforced to terminate training if the validation performance fails to improve for 15 consecutive epochs. To ensure a rigorous and fair comparison, our experimental configurations strictly align with the published settings of the baseline models, including the number of latent factors (K) and the architectural depth of the mapping networks.

In accordance with our theoretical framework, we formally characterize the dynamic market graph topology as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with a corresponding adjacency matrix A for the SGCN module. Node attributes are derived by mapping the Standard Industrial Classification (SIC) codes into the FF49 industry categories, which assigns a discrete sectoral label to each stock $v_i \in \mathcal{V}$. The edge set $\mathcal{E}$ is constructed based on a deterministic homophily constraint: an unweighted binary edge e_{ij} is established (where $A_{ij} = 1$) if and only if nodes i and j share an identical FF49 industry designation at a given timestep.

[1] https://pytorch.org/

Table A2: Statistical significance of architectural components for HeterIPCA across different numbers of latent factors (K). Diebold-Mariano test statistics comparing the out-of-sample MSE of the full HeterIPCA model against its three ablated versions (corresponding to Tests 1∼3) are reported. *** indicates a statistical significance at the 1% level.

	Test 1: α Network			**Test 2: SGCN Inter-firm Effect**			**Test 3: Dynamic Factor Encoding**		
K **Factors**	**ΔLoss** ($\times 10^{-4}$)	**NW-Z**	p**-value**	**ΔLoss** ($\times 10^{-4}$)	**NW-Z**	p**-value**	**ΔLoss** ($\times 10^{-3}$)	**NW-Z**	p**-value**
$K = 1$	1.46	5.42	< 0.001 ***	0.55	6.07	< 0.001 ***	3.52	7.57	< 0.001 ***
$K = 2$	1.04	5.09	< 0.001 ***	0.73	5.14	< 0.001 ***	3.78	7.81	< 0.001 ***
$K = 3$	1.59	5.79	< 0.001 ***	0.38	3.92	< 0.001 ***	4.32	7.52	< 0.001 ***
$K = 4$	0.54	6.97	< 0.001 ***	1.24	5.15	< 0.001 ***	3.63	8.28	< 0.001 ***
$K = 5$	1.45	6.32	< 0.001 ***	0.71	6.57	< 0.001 ***	4.11	7.82	< 0.001 ***
$K = 6$	1.68	7.56	< 0.001 ***	0.79	6.84	< 0.001 ***	4.06	7.88	< 0.001 ***

A.5 Statistical Significance

To validate the performance gains attributed to the inductive biases in HeterIPCA, we conduct pairwise Diebold–Mariano tests comparing out-of-sample Mean Squared Errors (MSE) between the full HeterIPCA and three primary ablation configurations:

- **Test 1 (without α-Net)**: Discards the firm-specific residual head.
- **Test 2 (without SGCN)**: Disables cross-firm message passing by replacing the empirical adjacency matrix with an identity matrix.
- **Test 3 (without Dyn-Factor)**: Substitutes the dynamic factor encoder with a static unconditional mean.

The tests evaluate the null hypothesis (H_0) of equal predictive accuracy against the one-sided alternative (H_1) that the full HeterIPCA is superior. To address heteroskedasticity and autocorrelation inherent in daily return data, we employ Newey-West HAC-adjusted Z-scores (NW-Z) with Andrews optimal truncation lags (Andrews, 1991). Results are evaluated across different numbers of latent factors $K \in \{1, \ldots, 6\}$ to ensure that the findings are not artifacts of specific model capacities, provided in Table A2.

Test 1: α-Net and Non-linear Mispricing
Ablating the α-Net incurs significant loss increases across all tested values of K (NW-Z $\in [5.09, 7.56]$, $p < 0.001$). The persistence of this effect in both low-capacity ($K = 1$) and high-capacity ($K = 6$) modes indicates that the residual signal is orthogonal to the linear factor span rather than a byproduct of insufficient capacity. Consequently, while the K-factor backbone captures systematic co-movement, the α-Net is essential for absorbing firm-level non-linear deviations that exceed the expressiveness of linear loadings.

Test 2: Graph Structure and Inter-firm Linkages
The removal of SGCN message passing significantly degrades predictive performance across all tested values of K (NW-Z > 3.9, $p < 0.001$). It confirms that inter-firm relationships captured through the graph structure provide distinct pricing signals that are fundamentally lost when relying only on individual firm features. This finding suggests that asset prices are influenced not only by idiosyncratic characteristics but also by their relational position within the broader economic network. Therefore, the proposed SGCN provides information that is essentially complementary to the traditional factor space, which validates the necessity of incorporating heterogeneous inter-firm linkages in modern IPCA frameworks.

Test 3: Influence of Dynamic Factor Encoding
Replacing the dynamic factor module with static means yields the largest effect size of all ablations. The resulting ΔLoss reaches 10^{-3} with NW-Z scores consistently exceeding 7.5 for all K, which has an order of magnitude greater than other components (10^{-4}). These results strongly reject the static baseline and highlight that cross-sectional pricing variance is heavily influenced by time-varying macro conditions. Without dynamic encoding, the ablated version fails to track regime shifts, which leads to compounded prediction errors.

To summarize, all three main modules of HeterIPCA provide significant performance improvements, with dynamic factor encoding identified as the most critical contributor. The stability of these results

across various K configurations suggests that the enhancements are a direct result of architectural design rather than hyperparameter tuning.

A.6 Turnover Analysis

Table A3: Average monthly turnover rates (%) of long-short and long-only portfolios across different numbers of latent factors (K). The turnover is calculated as the average absolute change in portfolio weights across consecutive monthly rebalancing periods.

Model	Long-short Portfolios Turnover (%)						Long-only Portfolios Turnover (%)					
	K=1	K=2	K=3	K=4	K=5	K=6	K=1	K=2	K=3	K=4	K=5	K=6
PCA	57.43	67.33	65.96	67.25	69.16	68.71	6.36	6.82	7.30	7.88	9.27	9.39
IPCA	73.76	73.35	64.24	79.66	111.95	124.95	35.38	37.63	31.73	38.90	54.33	60.45
CA0	65.59	77.92	130.98	142.29	143.64	142.11	35.65	41.97	66.53	70.86	71.53	70.99
CA1	62.57	57.19	63.38	64.22	68.06	69.33	35.01	33.27	34.27	34.87	36.08	36.29
CA2	60.71	58.70	57.18	57.65	56.15	55.92	34.81	34.78	34.20	34.86	34.01	33.24
CA3	62.64	62.70	61.61	59.47	61.76	61.33	34.64	33.37	33.12	33.69	33.75	34.52
SDNN	113.15	112.68	111.79	112.14	108.78	108.52	50.36	50.78	50.57	51.14	49.23	49.14
FactorGCL	90.37	125.83	124.96	124.49	124.26	122.25	45.82	58.58	57.82	57.03	56.91	56.18
CVAB	111.85	125.6	125.95	127.12	128.08	127.43	57.08	61.74	63.19	62.39	62.63	62.94
HeterIPCA (Ours)	123.81	122.16	122.70	124.12	123.92	124.37	58.24	58.31	57.93	58.55	58.13	58.21

To assess the practical implementation viability and the impact of transactional friction, we report the average monthly turnover rates for both long-short and long-only portfolios in Table A3. The turnover rate is defined as the average absolute percentage change in portfolio weights across consecutive monthly rebalancing periods. The results reveal several key findings.

First, deep learning and graph-based frameworks such as HeterIPCA, CVAB, SDNN, and FactorGCL yield higher turnover rates than linear baselines like PCA. For instance, the long-short turnover for HeterIPCA ranges between 122.16% and 124.37%, while PCA maintains a turnover below 70%. This elevated trading activity is expected: neural networks and graph convolutions are specifically designed to capture rapidly shifting and non-linear structural signals, such as dynamic peer effects. Exploiting these short-term patterns requires more frequent portfolio rebalancing.

Second, despite the higher baseline turnover, HeterIPCA shows strong stability across different numbers of latent factors (K). In contrast, traditional models like IPCA and CA0 experience significant turnover inflation as K increases, with rates rising to 124.95% and 143.64%, respectively. HeterIPCA maintains a steady turnover profile regardless of the dimensionality increase. This consistency suggests that HeterIPCA effectively captures underlying structural regularities rather than overfitting to high-frequency noise as the model complexity grows.

Third and most importantly, the higher turnover of HeterIPCA does not diminish its economic efficacy. As shown in the main results (Table 2), the out-of-sample Sharpe ratios are calculated net of a conservative 30 bps transaction cost per trade. HeterIPCA still achieves superior net Sharpe ratios (up to 2.57 for long-short portfolios), which confirms that the predictive signals extracted by the proposed SGCN architecture are sufficiently robust to overcome the associated transactional friction.

A.7 SHAP Analysis Across Market States

To characterize the state-dependent nature of conditional predictions from HeterIPCA, we compute Aumann-Shapley (SHAP) attributions across four cross-sections corresponding to archetypal market states in Figure A1.

(a) December 31, 1999: The peak of the dot-com valuation bubble.

(b) August 31, 2007: The quantitative equity squeeze (i.e., the "Quant Quake").

(c) October 31, 2008: The nadir of the Global Financial Crisis.

(d) March 31, 2020: The COVID-19 liquidity shock.

Features are ranked by mean absolute SHAP value, with color-coding indicating characteristic intensity (blue for low vs. red for high). Our analysis yields three primary insights into the structural behavior of HeterIPCA.

1. Risk Attribution Inversion

Risk-related characteristics show a systematic attribution inversion between expansionary and distressed states, specifically `retvol` (return volatility) and `beta`.

- **During the dot-com peak:** High-retvol and high-beta observations show strong positive SHAP values (up to +0.04), which suggests a market preference for systematic risk exposure.
- **During the 2008 and 2020 crises:** These color gradients reverse entirely. High-retvol and high-beta firms receive negative SHAP values (down to -0.03), while their low-value counterparts contribute positively.

This shift effectively captures the dynamic sign-switching of conditional risk prices between euphoric and stressed environments.

2. Regime-Specific Liquidity Premiums

The liquidity block reveals highly specific patterns, which comprises `zerotrade`, `baspread`, `ill`, and `dolvol`.

- **General Crisis Behavior:** During the 2008 crisis trough (Fig. A1c), illiquid stocks receive uniformly positive SHAP values, consistent with an activated illiquidity premium.
- **The 2007 Fingerprint:** The August 2007 panel (Fig. A1b) displays a unique inversion within this block. High-baspread and high-ill names are systematically penalized while large and liquid stocks are rewarded. This attribution pattern serves as a structural fingerprint of the crowded-quant deleveraging event, where liquid assets are sold to meet margin calls.

3. Momentum Horizon-Dependence

Momentum effects show significant horizon-dependence across regimes.

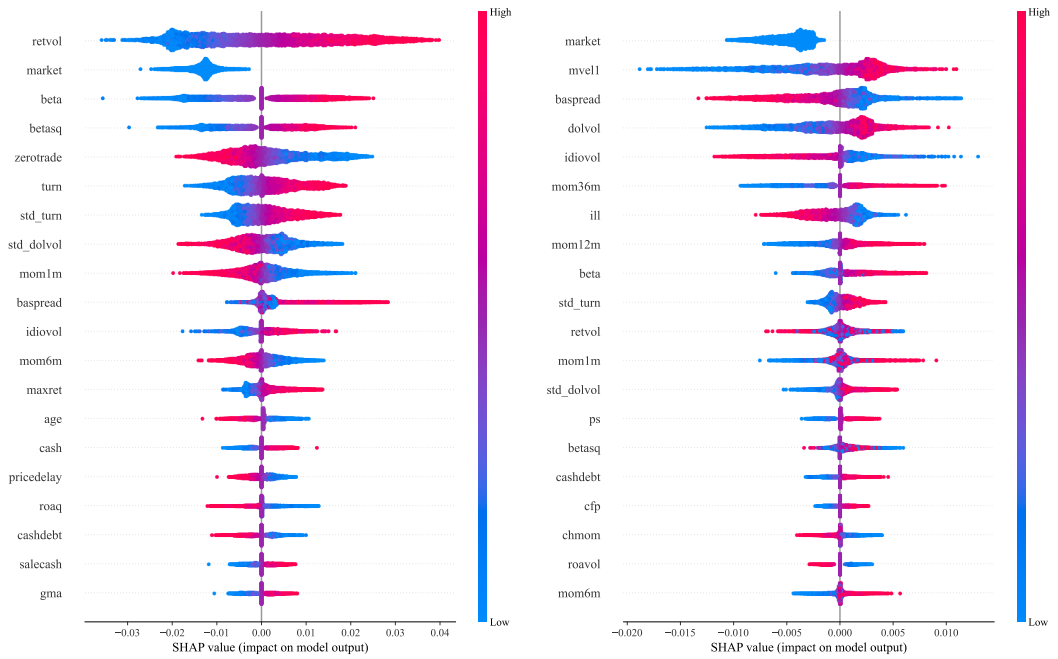
- **Short-term momentum (mom1m):** This transitions from a continuation signal in 1999 to a sharp reversal signal in 2008 and 2020.
- **Long-horizon signals (mom12m, mom36m):** These maintain a more stable continuation orientation throughout the sample, though their influence is largely attenuated during the COVID-19 shock.

To summarize, the SHAP geometry highlights two distinct forms of state sensitivity: substantial shifts in feature importance (such as `mvel1` rising to a top predictor in 2007) and clear sign-reversals in the transmission of risk and momentum. These findings validate the capacity of HeterIPCA to capture the non-linear and state-dependent relationships, and complement the aggregate results reported in the main text.

A.8 Ablation Analysis on Graph Structures

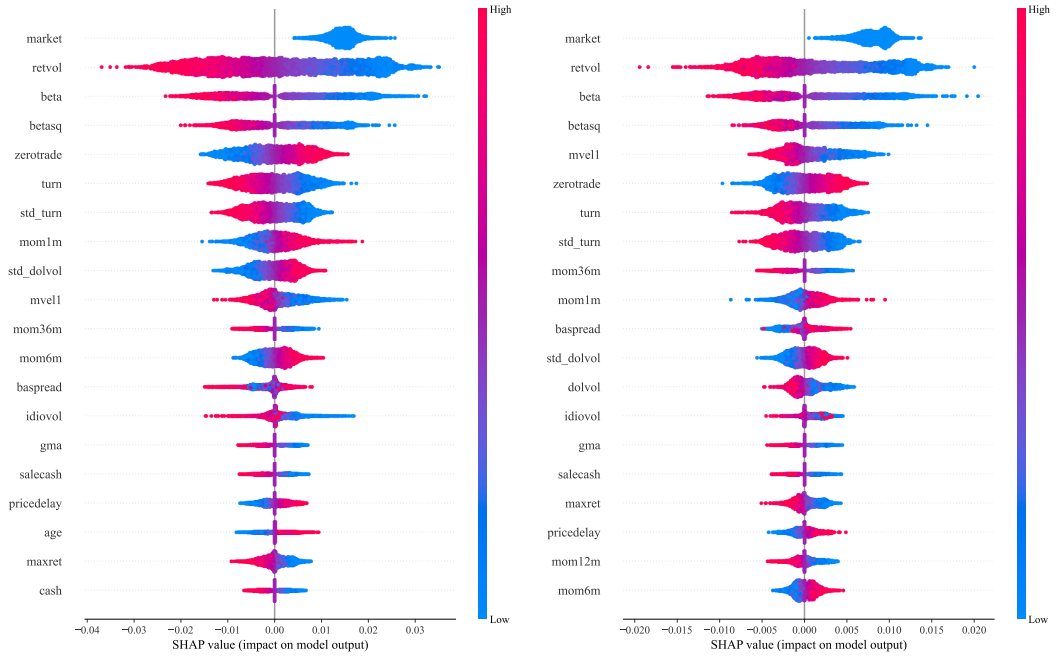
In the main experiments, the asset graph $\tilde{A} = A + \mathbb{I}$ is constructed by a standard financial domain structure based on industry sector co-membership (e.g., Standard Industrial Classification codes, SIC). Two assets i and j share an edge ($A_{ij} = 1$) if they belong to the same industry group, which reflects the well-established economic intuition that firms in the same sector share common business cycles, information channels, and risk exposures.

We further add 4 graph-construction ablations based on different edge rules: (1) a two-digit-SIC graph connecting stocks in the same two-digit SIC category; (2) a symmetric cosine-kNN graph with $k = 10$, constructed from the 94 contemporaneous rank-normalized firm characteristics; (3) a self-only graph without cross-stock edges; and (4) an Erdös-Rényi (ER) random graph whose expected monthly edge density is matched to that of the FF49 graph. The ER graph serves as a random-structure control to test whether the performance gain comes from the specific industry connections rather than simply from applying message passing over a graph with a similar density.



(a) Dot-Com Bubble (Dec 1999)

(b) Quant Quake (Aug 2007)



(c) Financial Crisis (Oct 2008)

(d) COVID-19 Crash (Mar 2020)

Figure A1: **Dynamics of Feature Importance Across Distinct Market Regimes.** We compare the Aumann-Shapley values (SHAP) for HeterIPCA across four iconic historical periods. Results show its ability to dynamically re-weight characteristics: (a) the 1999 tech-driven valuation bubble, (b) the 2007 systemic factor deleveraging, (c) the 2008 Global Financial Crisis, and (d) the 2020 pandemic-induced liquidity shock.

For each setting, we retrain HeterIPCA from scratch by changing only the edge set $\mathcal{E}_t$, while keeping the nodes, input features, model architecture, training procedure, and rolling evaluation windows unchanged.

The complete results for different factor numbers $K = 1, \ldots, 6$ are shown as Table A4. **FF49 (default)** achieves the **highest mean** under **both metrics**, while the two-digit SIC graph ranks second and is the closest alternative. This ranking indicates that the predictive gains of HeterIPCA persist across industry taxonomies. These comparisons indicate that the improvement does not arise from adding arbitrary edges alone. Economically meaningful cross-firm connections provide a more useful inductive bias than characteristic-based similarity or random connectivity.

Table A4: Ablation analysis for HeterIPCA on different graph structures.

Graph Structure	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	**Mean**
			R^2_r (%)				
FF49 (default)	0.482	**0.558**	0.562	**0.643**	**0.646**	**0.664**	**0.593**
SIC2	0.475	0.541	0.550	0.531	0.563	0.612	0.545
Cosine-kNN ($k=10$)	0.465	0.413	0.477	0.438	0.438	0.443	0.446
Self-only	**0.483**	0.392	**0.567**	0.500	0.533	0.603	0.513
Erdös-Rényi (ER)	0.401	0.428	0.444	0.462	0.471	0.487	0.449
			R^2_{rx} (%)				
FF49 (default)	0.473	**0.931**	1.095	**1.318**	**1.583**	**1.560**	**1.160**
SIC2	0.451	0.812	1.108	0.842	1.151	1.401	0.961
Cosine-kNN ($k=10$)	0.403	0.155	0.474	0.244	0.224	0.445	0.324
Self-only	**0.617**	0.375	**1.187**	0.785	1.138	1.378	0.913
Erdös-Rényi (ER)	0.322	0.298	0.521	0.603	0.742	0.864	0.558

A.9 Characteristic-level Interpretability Analysis

We add a characteristic-level interpretability analysis comparing different methods across $K = 1, \ldots, 6$. The Top-10 characteristic-alignment share is used to evaluate the stock-loading subspace. Results in Table A5 indicate that **HeterIPCA significantly outperforms other baselines in the Top-**10 **share alignment**, while keeping a relatively low cross-K stability.

Figure A2 further provides a practical interpretation of HeterIPCA at both the loading-subspace and transformed-axis levels. Figure A2(a) shows that the most strongly aligned characteristics at $K = 6$ are: `ill`, 0.765 [0.761, 0.772], `dolvol`, 0.745 [0.739, 0.755], `mvel1`, 0.743 [0.738, 0.747], `baspread`, 0.716 [0.712, 0.726], `retvol`, 0.706 [0.702, 0.721], `idiovol`, 0.700 [0.683, 0.719], `maxret`, 0.583 [0.566, 0.613], `zerotrade`, 0.568 [0.552, 0.575], `turn`, 0.535 [0.519, 0.545], and `betasq`, 0.414 [0.394, 0.428]. These characteristics primarily capture liquidity, trading frictions, firm size, trading activity, volatility, tail-risk, and market-exposure channels. Figure A2(b) shows that the adjacent-month Jaccard similarity has a median of 1.000, with identical Top-10 membership in 8 of the 11 transitions. This result indicates high short-horizon persistence of the characteristic footprint within the 2021 evaluation window. Under the fixed descriptive labeling transformation in Figure A2(c), the first two axes remain strongly associated with market exposure (the correlation $r = 0.80$) and illiquidity ($r = 0.88$) during 2009∼2021. The third axis has a weaker value-related association ($r = 0.33$), while the remaining axes are unresolved under the $|r| \geqslant 0.30$ criterion. We therefore interpret HeterIPCA mainly through its rotation-invariant loading subspace and characteristic groups, while treating individual axis labels as descriptive and transformation-dependent.

A.10 Practical Choice of Number of Latent Factors K

In empirical finance and factor modeling (Gu et al., 2021), $K \in \{1, \ldots, 6\}$ represents the standard baseline range for the number of factors. Empirical equity markets are generally dominated by a small set of pervasive macro/systematic risk factors. As shown in our experimental results (Tables 1 and 2), increasing K beyond 5 or 6 yields decreasing marginal returns in R^2_r and out-of-sample

Table A5: Characteristic-level interpretability analysis of differnet methods. [IQR] means the InterQuartile Range, where the bracketed values represent the 25^{th} percentile and 75^{th} percentile of the distribution across evaluation periods or iterations, and the main value is the median (50^{th} percentile). [block 95%] means the 95% Block Bootstrap Confidence Interval, where the bracketed values `[lower bound, upper bound]` form a 95% confidence interval estimated by a block bootstrap method.

Model	Top-10 share median ↑ [IQR]	Cross-K instability median ↓ [IQR]	Top-10 share mean ↑ [block 95%]	Paired mean: HeterIPCA−model [block 95%]	HeterIPCA higher months
HeterIPCA (ours)	**0.5002 [0.4819, 0.5025]**	0.1819 [0.1787, 0.1854]	**0.4953 [0.4828, 0.5094]**	**Reference**	**–**
CA3	0.4826 [0.4746, 0.4906]	0.4677 [0.4582, 0.4729]	0.4871 [0.4762, 0.5027]	0.0083 [0.0032, 0.0153]	10/12
FactorGCN	0.4621 [0.4544, 0.4657]	0.2624 [0.2554, 0.2728]	0.4630 [0.4530, 0.4731]	0.0324 [0.0292, 0.0374]	12/12
CA2	0.4587 [0.4550, 0.4634]	0.2595 [0.2578, 0.2645]	0.4607 [0.4529, 0.4699]	0.0346 [0.0276, 0.0423]	12/12
CA1	0.4176 [0.4122, 0.4243]	0.1724 [0.1683, 0.1784]	0.4183 [0.4134, 0.4242]	0.0771 [0.0661, 0.0881]	12/12
SDNN	0.4160 [0.4116, 0.4196]	0.0925 [0.0899, 0.0962]	0.4179 [0.4120, 0.4270]	0.0774 [0.0702, 0.0840]	12/12
CA0	0.4001 [0.3994, 0.4078]	**0.0197 [0.0191, 0.0204]**	0.4036 [0.3988, 0.4105]	0.0917 [0.0815, 0.1020]	12/12
IPCA	0.3978 [0.3958, 0.4026]	0.3570 [0.3545, 0.3582]	0.4002 [0.3965, 0.4070]	0.0951 [0.0843, 0.1048]	12/12
CAVB	0.3893 [0.3839, 0.3962]	0.2359 [0.2335, 0.2381]	0.3907 [0.3853, 0.3963]	0.1047 [0.0897, 0.1189]	12/12
PCA (reference)	0.3303 [0.3219, 0.3357]	0.0000 [0.0000, 0.0000]	0.3290 [0.3240, 0.3334]	0.1663 [0.1505, 0.1826]	12/12

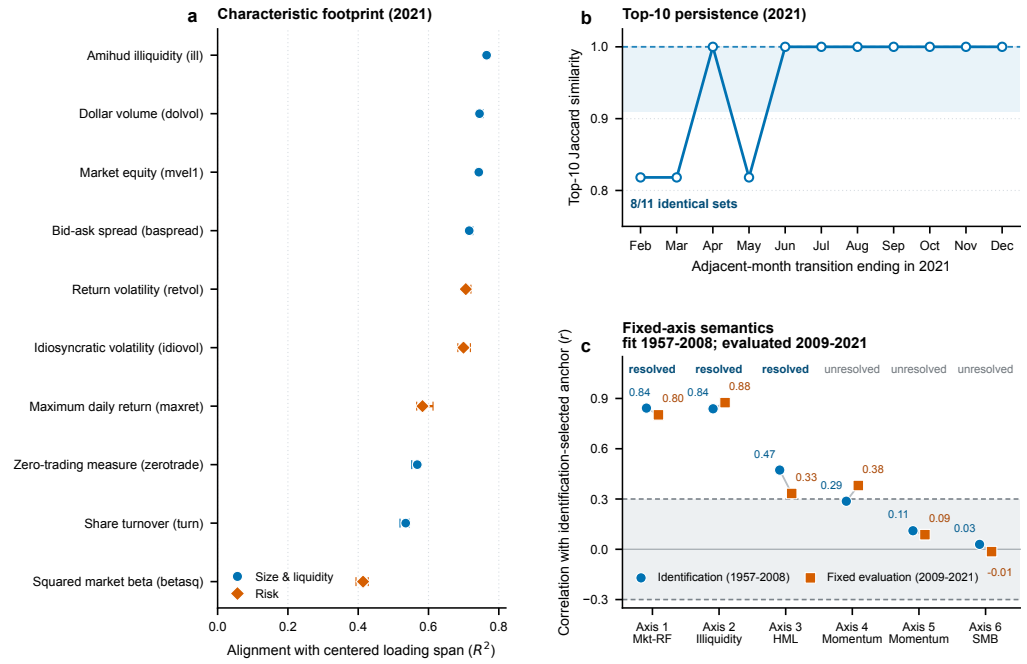


Figure A2: Practical interpretation of the HeterIPCA loading subspace at $K = 6$. **(a)** The ten characteristics with the largest rotation-invariant alignment with the centered loading subspace, ranked by their median R^2 across the 12 monthly cross-sections in 2021. Points denote medians, and horizontal bars denote interquartile ranges. Blue circles represent size- and liquidity-related characteristics, while orange diamonds represent risk-related characteristics. **(b)** Jaccard similarity between the month-specific Top-10 characteristic sets for the 11 adjacent-month transitions in 2021. The dashed line and shaded band denote the median and interquartile range, respectively. 8 transitions have identical Top-10 sets, while the remaining 3 have a Jaccard similarity of 0.818. **(c)** Correlations between each transformed axis and the anchor selected during 1957∼2008 (624 months), together with the correlations obtained by applying the same transformation without refitting during 2009–2021 (156 months). Blue circles and orange squares denote the identification and fixed-evaluation periods, respectively. The shaded region denotes the correlation $|r| < 0.30$. An axis is considered resolved only when its identification-selected anchor satisfies $|r| \geqslant 0.30$ in both periods. These axis associations are descriptive and transformation-dependent rather than a model-level holdout evaluation.

Sharpe ratios, which indicates that 5 to 6 factors adequately capture the primary common variance in cross-sectional asset returns without overfitting. We also add experiments for HeterIPCA on different values of K in Table A6. It indicates that $K = 5$ or 6 is a suitable choice for HeterIPCA.

A.11 Theoretical Assumptions and Practical Settings

Assumptions 1, 2, and 3 have been widely used in the literature (Chernozhukov et al., 2022a; Farrell et al., 2021), which are solidly grounded in economics and finance.

- **Assumption 1 (Sufficient Curvature and Boundedness):** The squared loss function ℓ provides sufficient curvature around the optimal parameter θ^*. The boundedness assumption is supported by the preprocessing procedure commonly adopted in empirical asset pricing: firm characteristics are cross-sectionally rank normalized to control extreme observations and ensure stable estimation.
- **Assumption 2 (Bounded Smoothness and Hölder Regularity $\mathcal{W}^{p,\infty}$):** This assumption reflects the economic intuition that factor exposures should vary systematically with firm characteristics instead of changing arbitrarily for infinitesimally close firms. Besides, a

Table A6: Performance of HeterIPCA across different values of K.

K	Total R^2_{x} (%)	Total R^2_r (%)	Predictive R^2_{x} (%)	Predictive R^2_r (%)
1	52.457	9.648	0.473	0.482
2	59.811	10.494	0.930	0.558
3	69.561	11.434	1.094	0.562
4	68.855	11.262	1.318	0.643
5	70.171	11.457	**1.583**	0.646
6	**70.993**	**11.567**	1.580	**0.664**
8	66.729	10.922	0.703	0.512
10	67.490	11.139	1.137	0.565
12	68.303	11.240	0.720	0.555
16	66.486	10.952	0.284	0.478
18	68.340	11.205	0.933	0.569
20	66.885	10.998	1.209	0.571
24	69.306	11.419	1.452	0.646
30	66.632	11.063	1.241	0.596

small increase in the market cap of a firm should lead to a smooth change in its systematic risk exposure, rather than a step-function jump.

- **Assumption 3 (Moment Conditions and DML Regularity):** Real-world asset returns usually show heavy tails and excess kurtosis (violating Gaussian distribution). However, this assumption only requires finite $q > 4$ moments. Besides, standard financial pre-processing (winsorizing extreme return outliers) ensures that higher-order moments remain bounded, which satisfies DML regularity conditions.
- Theorem 1 proves that the pseudo-dimension bound scales as $\mathcal{O}(QL\log(Q\mathfrak{n}))$. This theoretical result can explain the empirical findings in Table 1. Unregularized deep architectures (e.g., SDNN) overfit as the number of latent factors K increases, which causes portfolio-level R^2_{x} to drop to -0.748% at $K = 6$. In contrast, SGCN acts as a structural regularizer by parameter-sharing across node neighborhoods, which constrains representation complexity and enables HeterIPCA to scale smoothly to higher factor dimensions ($R^2_{\text{x}} = 1.560\%$ at $K = 6$).

AI USE STATEMENT

In this work, we used generative AI tools only for grammar correction and minor language edits to improve readability. We have not used generative AI tools for other tasks. We have reviewed all AI-assisted work, and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.