Flow-Based Arbitrage Pricing Theory

Abstract

I propose an arbitrage-pricing approach to analyze how noise trading flows impact asset prices. This approach uses no-arbitrage conditions to determine the impact of these flows on the stochastic discount factor, and consequently, on the cross-section of asset prices. By generalizing classic arbitrage theory, I show that noisy flows impact asset prices through a few important factors. The resulting model features rich patterns of cross-asset substitution beyond traditional mean-variance price impact models and logit demand systems. My theory also addresses prevalent misconceptions regarding the aggregation of asset flows into factor flows and the definition of factor-level price multipliers.

Keywords: arbitrage, factor, flow, price impact, risk

JEL Codes: G11, G12

1 Introduction

This paper studies the impacts of uninformed noise trading flows on asset prices. This is an important issue given extensive empirical evidence showing that noise trading flows, such as retail investor transactions or institutional investors' mechanical trading, substantially influence pricing across individual assets and the aggregate market.[1] Theoretical models suggest that arbitrageurs, constrained by their risk-bearing capacity, provide liquidity to these flows. Consequently, uninformed noise trading alters arbitrageurs' positions in risky assets, so price impacts arise as arbitrageurs' compensation for bearing extra risks.

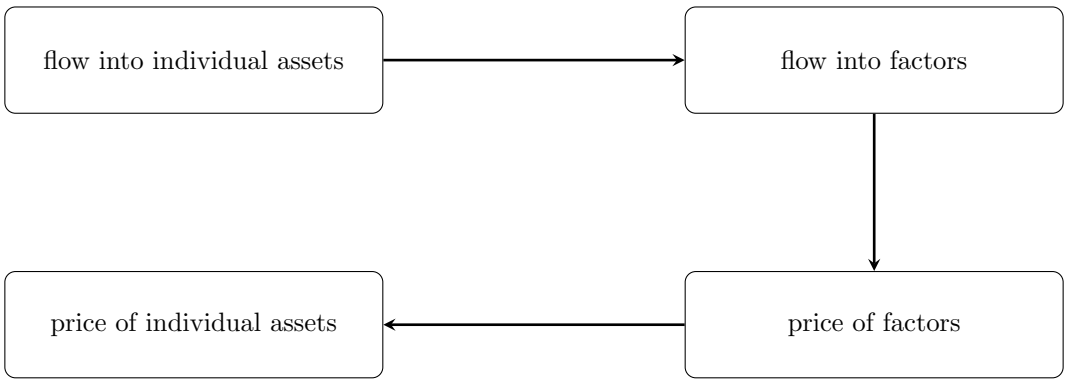
The literature has a gap in structurally estimating self and cross-asset price impacts. Typically, empiricists investigate noise trading flows into individual assets to quantify their price impacts. For example, suppose that uninformed retail investors buy \$100 million of Google and sell \$50 million of Facebook within a month. Estimating the price impacts requires understanding how Google's trading influences not only its own price but also Facebook's, and vice versa.[2] When extended to a large number of assets N, a reduced-form approach encounters the curse of dimensionality, requiring the estimation of an $N \times N$ matrix for self and cross-asset price impacts. This raises a critical question: how can one impose a theoretically grounded and empirically flexible structure on these $N \times N$ interactions?

This paper develops an arbitrage-pricing approach to answer this question. My theory builds upon the core asset-pricing equation $P = \mathbb{E}[MX]$, with P representing the asset price, M the stochastic discount factor (SDF), and X the asset payoff. The idea is that the impact of noisy flow on the asset price P must be reflected by its impact on the SDF M, because the flow is uninformed about the payoff X. Textbook theories offer various tools—portfolio theory, the Sharpe ratio, Hansen-Jagannathan bound, and Ross's arbitrage pricing theory

[1] See, for example, Coval and Stafford (2007), Lou (2012), Chang, Hong, and Liskovich (2015), Koijen and Yogo (2019), Huang, Song, and Xiang (2021), Barber, Huang, Odean, and Schwarz (2022), Ben-David, Li, Rossi, and Song (2022b), Gabaix and Koijen (2022), and Li and Lin (2022).

[2] Cross-asset price impacts are well-documented in studies like Boulatov, Hendershott, and Livdan (2013), Pasquariello and Vega (2015), and Chaudhary, Fu, and Li (2023).

Figure 1. The factor model of price impacts



Note: In the cross-section of assets, flows impacts prices through factors.

(APT)—to characterize how no-arbitrage conditions shape the SDF. This paper generalizes these tools to characterize how no-arbitrage conditions shape the SDF's response to noisy flows, thereby structuring the price impacts in the cross-section of assets. Unlike models limited to specific equilibria, my approach reveals the general cross-sectional constraints on price impacts in arbitrage-free markets. My approach avoids making parametric assumptions on utility function and payoff distribution, thereby improving upon equilibrium price impact models that rely on quadratic utility and normally distributed payoffs.[3]

My approach shows that noisy flows impact asset prices through a few important factors. (In this paper, "factors" refer to portfolios of assets.) The resulting factor model of price impacts unfolds in three steps, as illustrated in Figure 1. First, asset-level flows are aggregated into a few factors. Second, as flows impact prices by altering arbitrageurs' risk exposures, the price response of each factor to its own flow is quantified by the price change per unit of flow-induced risk. This measure improves upon the traditional price multiplier by incorporating risk exposures and avoiding mathematical mistakes when applying the traditional metric to long-short portfolios. Third, as factor prices change in response to flows, individual asset prices adjust accordingly to eliminate arbitrage. Under this model, rich patterns of

[3]See the survey article by Rostek and Yoon (2020). The typical assumptions made include a quadratic or CARA utility function and a multivariate normal payoff distribution. In many dynamic models with CRRA utility, e.g., Koijen and Yogo (2019), each investor's optimal portfolio problem effectively becomes static quadratic-normal after log-linearization.

cross-asset substitution emerge, beyond the traditional mean-variance price impact models and logit demand systems. In what follows, I present the three steps of the model.

The first step tackles a crucial yet unresolved question: How should individual asset flows be aggregated into flows into a given set of portfolios? This theoretical gap has led empirical studies to employ ad hoc methods for constructing portfolio flows, typically aggregating asset flows using portfolio weights, similar to calculating portfolio returns.[4]

I introduce a portfolio flow theory that refines flow aggregation in current literature. Recognizing that flows impact prices by altering arbitrageurs' risk exposures, a central principle of my theory emerges: keeping flow-induced risk exposures invariant when forming portfolios. Applying this principle, I show that portfolio flows equal the risk exposures (i.e., return betas) of the asset-level flows to the portfolios. This approach diverges from existing literature, which primarily relies on portfolio weights for aggregation. Under the specific assumption that each asset in the long leg of a long-short portfolio has a beta of 1 and each in the short leg has a beta of -1, traditional methods agree with my method.

Building on the portfolio flow theory, I proceed to the second step of the model: structuring the $N \times N$ matrix of price impacts for N assets through the lens of factors. This step requires taking a stance on the origins of cross-asset price impacts, for which I examine two mechanisms. The first, supported by empirical findings from Andrade, Chang, and Seasholes (2008) and Chaudhary, Fu, and Li (2023), is correlated payoffs. For instance, should noise traders purchase Google, arbitrageurs selling Google might buy Facebook to hedge against technology sector risk, creating a cross-impact. The second mechanism is correlated flows, backed by studies from Hasbrouck and Seppi (2001) and Ben-David, Li, Rossi, and Song (2022b), where noise traders simultaneously buying or selling assets like Google and Facebook induce cross-impacts.

I show that, irrespective of the flow and payoff correlation structures among the N assets, it is always possible to uniquely identify N factors characterized by both uncorrelated flows

[4]For a summary of empirical studies, see Table 1 in Gabaix and Koijen (2022).

and uncorrelated payoffs. Subsequently, I introduce the "Irrelevance of Uncorrelated Flows (IUF)" assumption, positing that these specific factors do not have cross-impacts on each other. To eliminate both channels of cross-impacts, it is pivotal to use the same set of factors to orthogonalize both flows and payoffs. This approach differs from prior studies, which analyze the factor structures of payoffs or flows in isolation.[5]

The IUF assumption simplifies the modeling of the $N \times N$ price impacts matrix to just N parameters. Each of these parameters quantifies a factor's price change per unit of risk induced by its own flow, termed "the price of flow-induced risk." For this parameter, explicitly measuring risk exposures is crucial, as flows impact prices by altering arbitrageurs' risk exposures. Traditionally, asset pricing models focus on the price of risk, defined as the ratio of expected returns to risk exposures. The newly introduced price of flow-induced risk represents how sensitive this traditional price of risk is to the flow for each factor.

Compared to the traditional price multiplier $(\Delta P/P)/(\Delta Q/Q)$, the price of flow-induced risk not only incorporates risk exposures but also avoids mathematical errors when the traditional metric is applied to long-short portfolios. For example, citing a series of empirical studies, Table 1 of Gabaix and Koijen (2022) claims a factor-level price multiplier of approximately 5, meaning that if noise traders buy 1% more of a factor, the factor price increases by 5%. However, the correct portfolio theory (which differs from the formula that this literature uses) implies that the total market capitalization Q of long-short factor portfolios is zero. This is because the market portfolio takes out the aggregate supply, so other long-short portfolios are net zero.[6] This results in a division by zero in the traditional formula, a clear mathematical error. Thus, the concept of "buying 1% more of a long-short portfolio" is fundamentally flawed, making any price multiplier estimate meaningless. It is erroneous to mechanically extend the formula $(\Delta P/P)/(\Delta Q/Q)$, which is well-defined for individual

[5] Ross (1976) models asset payoff and return factor structures. Lo and Wang (2000) and Hasbrouck and Seppi (2001) study the factor structures of trading volume and order flow. Gârleanu and Pedersen (2022) study the factor structures of both payoffs and flows, yet treat them separately in the analysis.

[6] The total amount outstanding of a long-short portfolio is not simply the difference between the sum of the long and short legs. Refer to Section 4.3 for a detailed derivation.

assets or aggregate markets, to long-short portfolios where it breaks down.

Equilibrium price impact models typically rely on quadratic utility and normally distributed payoffs. I show that the quadratic-normal model is a special case of my IUF model, but with the extra restriction that all factors must have the same price of flow-induced risk. This restriction arises because in typical equilibrium models, which considers a static setting with representative arbitrageurs, the price of flow-induced risk equals the ratio of risk aversion to the number of arbitrageurs—uniform across all factors.[7] Empirical evidence challenges the idea of a uniform price of flow-induced risk across different factors. Using mutual fund flows, An, Su, and Wang (2023) reject this uniformity restriction for the Fama and French (1993) three factors. In essence, standard quadratic-normal models have always implicitly assumed my IUF without realizing it, while inadvertently bundling in a less defensible restriction—a uniform price of flow-induced risk for all factors. By discarding this flawed restriction, the IUF allows for greater empirical flexibility and richer cross-substitution patterns.

Simply put, the IUF assumes no cross-impacts among some factors, and is special because it picks those factors with uncorrelated flows and uncorrelated payoffs. While this selection is empirically grounded in two cross-impact channels, a deeper theoretical justification for the optimality of IUF stands independently of any specific channel. I prove that, out of all possible ways of selecting N factors and eliminating cross-impacts among them, the IUF is the only way that keeps each factor's price response to its flow unchanged and keeps the arbitrageur's expected utility unchanged for any specific utility function.

The model's last step reduces the number of factors from N to a small K. The motivation is that in practice, N is usually large, and empirical researchers often use a small set of factors to achieve robustness in estimation. Doing so requires generalizing the concept of arbitrage to markets with noise trading. Traditionally, no arbitrage means that a portfolio with little payoff risk should not have a high expected return. I generalize this principle for noise

[7] To relax this restriction in equilibrium models, two approaches are viable: firstly, introducing predictable flows within a dynamic model, as detailed in Section 7.1; secondly, allowing for distinct arbitrageur groups for different factors, as detailed in An, Su, and Wang (2023).

trading scenarios, postulating that a small flow into a portfolio with little payoff risk should not generate a high price impact.[8] If such portfolios were to exist, arbitrageurs would earn exceptionally favorable risk-return trade-offs for providing liquidity to these portfolios.

Equipped with the generalized concept of arbitrage, I proceed to develop the flow-based APT. I prove that if the K factors' flows explain common variations in the N assets' flows, then the N assets' price impacts must align with the K factors' price impacts in accordance with the assets' risk exposure to these factors.[9] If this were not the case, one could identify an asset, hedge out the factors, and end up with a portfolio with a small flow and little risk, yet a significant price impact, contradicting the generalized concept of arbitrage. Importantly, my theory does not rely on the assumption that noise traders understand factors or covariances. Regardless of the underlying behavioral models of why noise traders generate flows, it is the arbitrage force that aligns the price impacts of these flows with the modeled factor structure.

From an application perspective, my theory expands the class of tractable asset-pricing models with trading flows. Provided a single-asset model is solvable, researchers can apply my machinery to solve the general model with N assets featuring any correlation pattern in both payoffs and flows. The existing literature's need to impose stringent correlation structures for tractability is now obsolete. This advancement not only facilitates theoretical investigations into how trading flows impact cross-sectional asset prices in a more realistic setting, but also provides a unified method for empirical estimation of such effects. To illustrate, the paper presents two applications from subsequent works where flows either exhibit dynamic predictability or are informed about asset payoffs.[10]

The prediction of the model—that noisy flows impact cross-sectional asset prices through

[8] In my theory, arbitrage still holds: two assets with identical payoffs should have the same price, even though that same price can be affected by noise trading. My theory does not account for "noise trader risks," where two assets with identical payoffs can have different prices if noise trading itself becomes an independent risk factor, as in De Long, Shleifer, Summers, and Waldmann (1990).

[9] Order flows indeed exhibit a pronounced factor structure, as empirically documented by Hasbrouck and Seppi (2001). Their finding is motivated using optimization models of trading (Caballe and Krishnan, 1994; Kumar and Seppi, 1994).

[10] In the microstructure literature, a technical trick involves rotating assets to portfolios with uncorrelated payoffs, subsequently assuming that signals on these uncorrelated payoffs are also uncorrelated (see chapter 3.8 of Veldkamp, 2011). Section 7.2 explains the advantages of my rotation over the traditional approach.

a few important factors—finds empirical support in An, Su, and Wang (2023). They estimate the model using mutual fund flows data, applying it to the Fama-French 5×5 size and book-to-market test assets, alongside the Fama-French three factors. Their results affirm that the model, which relies solely on the estimated price of flow-induced risk of the three factors, effectively explains the observed patterns of self- and cross-impacts among the 5×5 assets.

The remainder of this paper is organized as follows. Section 2 reviews related literature. Section 3 presents the model setup. Sections 4, 5, and 6 present the three steps of the model. Section 7 presents two applications. Section 8 concludes. The appendices provide proofs.

2 Related Literature

Cross-sectional asset pricing traditionally builds on factor models (Fama and MacBeth, 1973; Merton, 1973a; Ross, 1976), while demand-based studies highlight how uninformed noise trading flows can significantly impact asset prices.[11] Although the significance of both factors and demand in asset pricing is well-recognized, a unifying theoretical framework is absent. My contribution is to fill this gap by generalizing classic factor-model theories to analyze markets affected by noise trading. This generalization yields a new factor model that provides a theoretically grounded and empirically flexible structure for the $N \times N$ self- and cross-impacts. Importantly, existing literature, such as summarized by Table 1 of Gabaix and Koijen (2022), makes mistakes on two key issues: 1) the construction of portfolio flows, and 2) the definition of price multipliers for long-short portfolios. My contribution is identifying and correcting these mistakes.

Arbitrage theory centers around $P = \mathbb{E}[MX]$, which underpins portfolio theory, the Sharpe ratio, Hansen-Jagannathan bound, and Ross's APT. These tools characterize how no-arbitrage conditions shape the SDF and, consequently, determine cross-sectional asset

[11] See, for example, Coval and Stafford (2007), Lou (2012), Boulatov, Hendershott, and Livdan (2013), Chang, Hong, and Liskovich (2015), Pasquariello and Vega (2015), Koijen and Yogo (2019), Huang, Song, and Xiang (2021), Barber, Huang, Odean, and Schwarz (2022), Ben-David, Li, Rossi, and Song (2022b), Gabaix and Koijen (2022), Li and Lin (2022), and Chaudhary, Fu, and Li (2023).

prices. My contribution is generalizing these tools to characterize how no-arbitrage conditions shape the SDF's response to noisy flows, thus structuring the cross-sectional price impacts. Unlike models limited to specific equilibria, my approach reveals the general cross-sectional constraints on price impacts in arbitrage-free markets, and avoids making parametric assumptions on utility function and payoff distribution. My theory complements Kozak, Nagel, and Santosh (2018), who show that arbitrage leads to a factor structure in asset returns in the presence of noise traders. Differing from their focus on asset returns, I study the price impacts of noisy flows and derive a factor model for these impacts.

The standard quadratic-normal model implicitly assumes that all factors must have the same price sensitivity per unit of flow-induced risk. My IUF improves over the standard model by allowing heterogeneous price of flow-induced risk among different factors. Consequently, the IUF model implies that the cross-substitution between two assets depends not only on their factor risk exposures but also on the distinct price of flow-induced risk specific to each factor. Hence, some cross-asset price impacts that do not show up in the quadratic-normal model emerge under my model. The intuition is that flow into asset A can impact the price of asset B by differentially impacting the factors that asset B is exposed to. The risk-based cross-substitution distinguishes my model from existing models, such as Koijen and Yogo (2019), which features proportional cross-substitution in the logit demand system, and Buffa and Hodor (2023), where cross-substitution is determined by whether assets are held in the same or different benchmarks.[12]

In my model, price impacts stem from arbitrageurs' aversion to absorbing flow-induced risk, not from microstructure or liquidity frictions. This feature distinguishes my model from studies on the effects of commonality in liquidity on expected returns (Chordia, Roll, and Subrahmanyam, 2000; Hasbrouck and Seppi, 2001; Pástor and Stambaugh, 2003). Unlike models where common flows are intrinsic risk factors (Alvarez and Atkeson, 2018; Dou, Kogan, and Wu, 2021; Kim, 2020), in my model, these flows alter the pricing of risk factors.

[12]Specifically, in Koijen and Yogo (2019), two assets experience positive cross-substitution if held by the same institution. In Buffa and Hodor (2023), heterogeneous benchmark holdings lead to negative spillovers.

3 Model Setup and Flow-Based SDF

In this section, I introduce the model setup and the flow-based SDF, a tool central to characterizing the cross-section of price impacts.

3.1 Model Setup

The model unfolds over two periods, $t = 0$ and $t = 1$. Define the probability space as $(\Omega, \mathcal{F}, \mathbb{P}) = (\Omega_0 \times \Omega_1, \mathcal{F}_0 \times \mathcal{F}_1, \mathbb{P}_0 \times \mathbb{P}_1)$. Let $\omega_0 \in \Omega_0$ and $\omega_1 \in \Omega_1$ represent the randomness at times $t = 0$ and $t = 1$, respectively. To ensure clarity, I explicitly specify the dependence of random variables on ω_0 and ω_1 when necessary. As a convention, I use bold font notation for matrices and vectors, and $\mathbf{A}^\top$ to denote the transpose of matrix $\mathbf{A}$.

The model includes N assets. The flow into asset $n = 1, \ldots, N$ is represented by $f_n(\omega_0)$, a random variable that realizes at time 0, grouped as vector $\mathbf{f} = (f_1, f_2, \ldots, f_N)^\top$. These flows are measured in dollar values, relative to the asset prices before any flow occurs. The payoff of asset n is denoted by $X_n(\omega_1)$, a random variable that realizes at time 1, grouped as vector $\mathbf{X} = (X_1, X_2, \ldots, X_N)^\top$. The flow is uninformed and noisy in the sense that $\mathbf{f}$ is independent of the payoff $\mathbf{X}$. The gross risk-free rate is a constant, R_F, and does not depend on the flow. Without loss of generality, I assume that $\text{var}(\mathbf{X})$ and $\text{var}(\mathbf{f})$ are full-rank matrices.[13] Notably, I do not assume any parametric distributions for flow or payoff.

I denote the time-0 price of asset n as $P_n(\mathbf{f})$, grouped as vector $\mathbf{P}(\mathbf{f}) = (P_1(\mathbf{f}), P_2(\mathbf{f}), \ldots, P_N(\mathbf{f}))^\top$. In contrast to the equilibrium approach, the arbitrage approach does not model specific economic agents. However, one can imagine a representative arbitrageur who absorbs the flow $\mathbf{f}$, allowing the price $\mathbf{P}(\mathbf{f})$ to depend on $\mathbf{f}$. Figure 2 displays the model timeline.

Note that the flow $\mathbf{f}$ arrives at time 0. There could, in principle, also be flows arriving between time 0 and 1, potentially affecting the "payoff" $\mathbf{X}$ at time 1. By assuming $\mathbf{f}$ is independent of $\mathbf{X}$, my analysis effectively conducts a comparative static exercise, varying

[13] Otherwise, one can select linearly independent portfolios and rotate the payoff $\mathbf{X}$ and flow $\mathbf{f}$ accordingly.

Figure 2. Model timeline

Notes: The probability space is defined as $(\Omega, \mathcal{F}, \mathbb{P}) = (\Omega_0 \times \Omega_1, \mathcal{F}_0 \times \mathcal{F}_1, \mathbb{P}_0 \times \mathbb{P}_1)$. The flow $\mathbf{f}$, asset price $\mathbf{P}(\mathbf{f})$, and price impact $\Delta\mathbf{p}(\mathbf{f})$ are random variables that realize at time 0. Asset payoff $\mathbf{X}$ and fundamental return $\mathbf{R}_0$ are random variables that realize at time 1. For any flow $\mathbf{f}$, the flow-based SDF $\Delta M(\mathbf{f})$ is a random variable that realizes at time 1.

the time-0 flow $\mathbf{f}$ while holding all else equal. There might be concerns that the time-0 flow $\mathbf{f}$ could dynamically predict future flows or that, in general, the flow $\mathbf{f}$ might be informed about the payoff $\mathbf{X}$. My theory can be generalized to accommodate both these possibilities, as discussed in Sections 7.1 and 7.2, respectively.

3.2 Flow-Based SDF

To characterize how flow impacts price for the cross-section of assets, I introduce the flow-based SDF. Intuitively, because the flow changes the asset price $\mathbf{P}(\mathbf{f})$ but does not change the future payoff $\mathbf{X}$, the SDF bridging asset price to payoff must consequently adjust. Specifically, the law of one price (LOOP) implies that for any given flow $\mathbf{f}$, there exists a standard SDF $M(\mathbf{f})$, which is a random variable that realizes at time 1, such that[14]

$$\mathbf{P}(\mathbf{f}) = \mathbb{E}[M(\mathbf{f})\mathbf{X}]. \tag{1}$$

Note that the expectation on the right-hand side is taken over time-1 random variables $M(\mathbf{f})$ and $\mathbf{X}$ for any given flow $\mathbf{f}$. When the flow $\mathbf{f}$ equals zero, I obtain

$$\mathbf{P}(\mathbf{0}) = \mathbb{E}[M(\mathbf{0})\mathbf{X}]. \tag{2}$$

[14]My theory requires only the LOOP, ensuring the existence of an SDF. I do not require the additional no-arbitrage condition to guarantee a strictly positive SDF (see chapter 4.2 of Cochrane, 2009).

In this context, whether the flow $\mathbf{f}$, asset price $\mathbf{P}(\mathbf{f})$, SDF $M(\mathbf{f})$, and payoff $\mathbf{X}$ are "exogenous" or "endogenous" is irrelevant. Regardless of their nature, equation (1) must hold in any equilibrium that satisfies LOOP. This arbitrage-pricing logic, which has been developed since the works of Black and Scholes (1973), Merton (1973b), and Ross (1976), enables the characterization of the price impact of flows across various market equilibria, extending beyond specific models.

For any flow $\mathbf{f}$, I define *price impact* as

$$\Delta\mathbf{p}(\mathbf{f}) := \left(\frac{P_1(\mathbf{f}) - P_1(\mathbf{0})}{P_1(\mathbf{0})}, \frac{P_2(\mathbf{f}) - P_2(\mathbf{0})}{P_2(\mathbf{0})}, \ldots, \frac{P_N(\mathbf{f}) - P_N(\mathbf{0})}{P_N(\mathbf{0})}\right)^\top, \tag{3}$$

which represents the percentage price change at time 0 with and without flow $\mathbf{f}$. I define *fundamental return* as

$$\mathbf{R}_0 := \left(\frac{X_1}{P_1(\mathbf{0})}, \frac{X_2}{P_2(\mathbf{0})}, \ldots, \frac{X_N}{P_N(\mathbf{0})}\right)^\top, \tag{4}$$

which represents asset return when there is no flow and is independent of the flow $\mathbf{f}$. I define the *flow-based SDF (F-SDF)* as the difference between two standard SDFs with and without flow $\mathbf{f}$,

$$\Delta M(\mathbf{f}) := M(\mathbf{f}) - M(\mathbf{0}), \tag{5}$$

which captures the flow-induced changes in the standard SDF.[15]

Taking the difference between (1) and (2) and dividing element-wise by $\mathbf{P}(\mathbf{0})$, I obtain

$$\Delta\mathbf{p}(\mathbf{f}) = \mathbb{E}[\Delta M(\mathbf{f})\mathbf{R}_0]. \tag{6}$$

Since the fundamental return $\mathbf{R}_0$ is independent of flow $\mathbf{f}$, equation (6) implies that flow generates price impact only by varying the F-SDF $\Delta M(\mathbf{f})$. The insight is that to characterize

[15]The term "flow-based SDF" was first introduced in section 5.3 of Gabaix and Koijen (2022), where they solve a general equilibrium model and express the model's solutions using an SDF that depends on flows. Many papers also presuppose certain functional forms for how flow influences the SDF. In contrast, my approach derives how flow changes the SDF through no-arbitrage conditions, independent of any particular equilibrium model. This new approach leads to the derivation of my F-SDF pricing equation (6), which forms the foundation for all my subsequent analyses.

the multi-asset price/quantity relationship $\Delta\mathbf{p}(\mathbf{f})$, it is only necessary to determine how the flow $\mathbf{f}$ changes the SDF $\Delta M(\mathbf{f}) = M(\mathbf{f}) - M(\mathbf{0})$, rather than tackling the more complex question of how the entire SDF $M(\mathbf{f})$ depends on $\mathbf{f}$. Economically, all the risk preferences of how the arbitrageur responds to flows are encoded by the F-SDF mapping,

$$\tilde{M}(\cdot): \quad \mathbf{f} \in \mathbb{R}^N \longrightarrow \tilde{M}(\mathbf{f}, \omega_1) \in L^2(\Omega_1, \mathcal{F}_1, \mathbb{P}_1), \tag{7}$$

where $L^2(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ is the set of square-integrable time-1 random variables. Mathematically, the mapping $\tilde{M}(\cdot)$ is called a random field. The technical challenge lies in tracking multi-asset flows and returns, as well as their relationship, while the standard arbitrage pricing tracks only returns. Therefore, I use a random field to characterize the F-SDF, similar to how a random variable is used to characterize the standard SDF.

The following proposition summarizes the key equations for the F-SDF.

PROPOSITION 1. *For any flow* $\mathbf{f}$*, the F-SDF* $\Delta M(\mathbf{f})$ *satisfies*

$$\mathbb{E}[\Delta M(\mathbf{f})\mathbf{R}_0] = \Delta\mathbf{p}(\mathbf{f}), \tag{8}$$

$$\mathbb{E}[\Delta M(\mathbf{f})] = 0. \tag{9}$$

Equation (8) simply restates (6), which is a consequence of the LOOP. Equation (9) implies that the F-SDF has zero mean and follows from the assumption that the risk-free rate R_F does not depend on the flow $\mathbf{f}$.

In summary, this section establishes the foundational linkages between the F-SDF $\Delta M(\mathbf{f})$ and the price impact $\Delta\mathbf{p}(\mathbf{f})$. In subsequent sections, I leverage a series of axiomatic principles to characterize the F-SDF, which in turn enables me to determine the price impacts.

4 Portfolio Theory for Flows

This section presents the first step of the model, which tackles a fundamental yet unresolved question: How should individual asset flows be aggregated into flows into a given set of portfolios? Section 4.1 establishes the foundational principle of the portfolio flow theory: the flow-induced risk exposures should remain invariant when forming portfolios. Section 4.2 applies this principle to construct portfolio flows. Section 4.3 compares my formula with those found in the literature, demonstrating how my formula both generalizes and corrects the existing approaches.

4.1 Invariance Principle

This section establishes the foundational principle of the portfolio flow theory: the overall risk exposures induced by flows should remain invariant under portfolio formation. To understand this, consider that when the arbitrageur absorbs flow $\mathbf{f}$, the change in wealth is given by[16]

$$\Delta W(\mathbf{f}) = R_F \mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f}) - \mathbf{f}^\top (\mathbf{R}_0 - R_F \mathbf{1}). \tag{10}$$

Here, the inner product between flow and price impact, $\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f})$, represents the total compensation to the arbitrageur for absorbing the risk induced by the flow. Meanwhile, the inner product between flow and excess fundamental return, $\mathbf{f}^\top (\mathbf{R}_0 - R_F \mathbf{1})$, represents the payoff risk absorbed by the arbitrageur.

Portfolio formation is simply a different way to look at the same problem. This implies that regardless of whether individual assets or aggregated portfolios are considered, the arbitrageur's wealth change $\Delta W(\mathbf{f})$ should stay invariant. According to equation (10), this requires keeping the inner product $\mathbf{f}^\top (\mathbf{R}_0 - R_F \mathbf{1} - R_F \Delta \mathbf{p}(\mathbf{f}))$ invariant under portfolio for-

[16] To derive this equation, note that the flow $\mathbf{f}$ is measured per dollar, which allows expressing the flow in units of shares as $\mathbf{h} = (f_1/P_1(\mathbf{0}), \ldots, f_N/P_N(\mathbf{0}))^\top$. The arbitrageur sells $\mathbf{h}$ units of shares at time 0 at price $\mathbf{P}(\mathbf{f})$. Consequently, $\Delta W(\mathbf{f}) = R_F \mathbf{h}^\top \mathbf{P}(\mathbf{f}) - \mathbf{h}^\top \mathbf{X}$, simplifying to (10) by using equations (3) and (4).

mation. The term $\mathbf{R}_0 - R_F\mathbf{1} - R_F\Delta\mathbf{p}(\mathbf{f})$ is approximately the excess fundamental return from the time-0 price $\mathbf{P}(\mathbf{f})$ to the time-1 payoff[17] $\mathbf{X}$. Denoting this excess return as $\mathbf{R}$, the principle of the portfolio flow theory is to keep the inner product $\mathbf{f}^\top\mathbf{R}$ invariant, which represents the overall risk exposures induced by flows.

4.2 Formula for Portfolio Flows

This section applies the invariance principle to construct portfolio flows. Suppose that one forms $K \leq N$ portfolios using N assets. I denote the $N \times K$ portfolio weight matrix as $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K)$, where $\mathbf{b}_k = (b_{1,k}, b_{2,k}, \ldots, b_{N,k})^\top$ is the k-th portfolio's weight, meaning that portfolio k holds $b_{n,k}$ dollars of asset n. The goal is to construct the flow $\mathbf{q} = (q_1, q_2, \ldots, q_K)^\top$ into the K portfolios. Because the number K of portfolios may be less than the number N of assets, it is unrealistic to expect a complete recovery of the flow-induced risk exposure $\mathbf{f}^\top\mathbf{R}$ using the K portfolios. Instead, a reasonable objective to minimize the variance of the difference between the N assets and K portfolios. Recall that standard portfolio theory implies that the returns of the K portfolios are $\mathbf{B}^\top\mathbf{R}$ (i.e., the return of portfolio k is $\sum_{n=1}^N b_{n,k}R_n$). Therefore, I have

$$\mathbf{q} = \arg\min \operatorname{var}(\mathbf{f}^\top\mathbf{R} - \mathbf{q}^\top\mathbf{B}^\top\mathbf{R}) = \operatorname{var}(\mathbf{B}^\top\mathbf{R})^{-1}\operatorname{cov}(\mathbf{B}^\top\mathbf{R}, \mathbf{f}^\top\mathbf{R}). \tag{11}$$

Equation (11) implies that the portfolio flows $q_1, q_2, \ldots, q_K$ result from projecting the flow-induced risk exposure $\mathbf{f}^\top\mathbf{R}$ onto the returns of K portfolios, $\mathbf{b}_1^\top\mathbf{R}, \mathbf{b}_2^\top\mathbf{R}, \ldots, \mathbf{b}_K^\top\mathbf{R}$. In other words, portfolio flows $\mathbf{q}$ are the return betas of the asset flows $\mathbf{f}$ to the portfolios $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_K$, which is equivalently the projection of $\mathbf{f}$ onto $\mathbf{B}$ within the space of return.[18]

[17] Using equations (3) and (4), we have $\mathbf{R}_0 - R_F\mathbf{1} - R_F\Delta\mathbf{p}(\mathbf{f}) = ((X_1 - R_F P_1(\mathbf{f}))/P_1(\mathbf{0}), \ldots, (X_N - R_F P_N(\mathbf{f}))/P_N(\mathbf{0}))^\top$. Here, the denominator is $P_n(\mathbf{0})$ instead of $P_n(\mathbf{f})$ because returns are measured relative to the asset prices before any flow occurs.

[18] An alternative approach is to directly project $\mathbf{f}$ onto $\mathbf{B}$ using the equation $\mathbf{q} = (\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{f}$, which does not need to use the return $\mathbf{R}$. While this method does not minimize the objective in (11), it maintains the projected risk exposure on the portfolios $\mathbf{B}$ as invariant. This is because $\mathbf{f}^\top\mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{R} = ((\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{f})^\top\mathbf{B}^\top\mathbf{R}$, with $\mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top$ being the projection matrix.

To help readers better understand equation (11), I write

$$\mathbf{q} = \text{var}(\mathbf{B}^\top \mathbf{R})^{-1}\text{cov}(\mathbf{B}^\top \mathbf{R}, \mathbf{f}^\top \mathbf{R}) = \sum_{n=1}^{N} f_n \text{var}(\mathbf{B}^\top \mathbf{R})^{-1}\text{cov}(\mathbf{B}^\top \mathbf{R}, R_n), \tag{12}$$

where $\text{var}(\mathbf{B}^\top \mathbf{R})^{-1}\text{cov}(\mathbf{B}^\top \mathbf{R}, R_n)$ is the return betas of asset n to the K portfolios. Equation (12) implies that the portfolio flows $\mathbf{q}$ are obtained by aggregating the asset flows f_n. Rather than simply accumulating these flows based on portfolio weights, they are weighted by each asset's risk exposures (i.e., return betas) to the portfolios. Economically, the portfolio flow q_k measures how the asset-level flow $\mathbf{f}$ changes risk exposures for portfolio k.

In the special case of forming $K = N$ portfolios using N assets, equation (11) simplifies to

$$\mathbf{q} = \mathbf{B}^{-1}\mathbf{f} \text{ or, equivalently, } \mathbf{f} = \mathbf{B}\mathbf{q} = \sum_{k=1}^{N} \mathbf{b}_k q_k, \tag{13}$$

which has a more intuitive interpretation. Portfolio flows q_k first enter portfolio k and are then allocated to individual assets based on portfolio weights $\mathbf{b}_k$. That is, if portfolio-k flow increases by \$1, asset-n flow rises by $\$b_{n,k}$. Therefore, if the N portfolio flows are $q_1, q_2, \ldots, q_N$, the resulting flow into asset n is $\sum_{k=1}^{N} b_{n,k} q_k$, which implies $\mathbf{f} = \sum_{k=1}^{N} \mathbf{b}_k q_k$. Consequently, to aggregate asset flows into portfolio flows, one should project asset flows onto portfolio weights, explaining the inverse $\mathbf{B}^{-1}$.

4.3 Comparison with the Literature's Formula

This section compares my formula (12) with the literature's formula, which directly sums asset flows based on portfolio weights. This comparison highlights the improvements of my formula over existing literature, particularly for long-short portfolios.

First, my formula implies that the flow into a portfolio can depend on the weights of other portfolios. As per (12), this is because the return beta to a portfolio can depend on whether other portfolios are controlled for, a standard intuition in asset pricing. This is in stark contrast to the literature's formula, where the flow into a portfolio solely depends

on its own portfolio weights. Nonetheless, in the special case that the K portfolios have uncorrelated returns (i.e., $\text{cov}(\mathbf{b}_k^\top \mathbf{R}, \mathbf{b}_j^\top \mathbf{R}) = 0$ for $k \neq j$), my formula can also be simplified to a portfolio-by-portfolio construction, i.e., $q_k = \text{var}(\mathbf{b}_k^\top \mathbf{R})^{-1} \text{cov}(\mathbf{b}_k^\top \mathbf{R}, \mathbf{f}^\top \mathbf{R})$.

Second, for the market portfolio, my formula coincides with the literature's formula under an additional assumption. I denote the market portfolio weights as $b_n^{\{\text{MKT}\}} = Q_n / \sum_{n=1}^N Q_n$, where Q_n is the market capitalization of asset n. The literature's formula aggregates the market-capitalization-normalized flow f_n/Q_n in accordance with the portfolio weight $b_n^{\{\text{MKT}\}}$,

$$\sum_{n=1}^N b_n^{\{\text{MKT}\}} \frac{f_n}{Q_n} = \frac{\sum_{n=1}^N f_n}{\sum_{n=1}^N Q_n}. \tag{14}$$

The numerator $\sum_{n=1}^N f_n$ measures the "aggregate flow into the market." It is conceptually different from $q^{\{\text{MKT}\}}$ that I measure, which is the "flow into the market portfolio." Under an extra assumption that asset-level flows are solely driven by the flow into market portfolio ($f_n = b_n^{\{\text{MKT}\}} q^{\{\text{MKT}\}}$ for all n), my measure $q^{\{\text{MKT}\}}$ equals the literature's measure $\sum_{n=1}^N f_n$.

Third, the literature (see Table 1 of Gabaix and Koijen (2022) for a summary) also extends formula (14) to measure flows into long-short portfolios, leading to a mistaken interpretation. Consider a long-short portfolio, in which the portfolio weight for the long leg is $b_n^{\{LS\}} = Q_n / \sum_{n\in\{L\}} Q_n$ if $n \in \{L\}$ and the portfolio weight for the short leg is $b_n^{\{LS\}} = -Q_n / \sum_{n\in\{S\}} Q_n$ if $n \in \{S\}$. Note that this construction aligns with the standard practice in long-short portfolios, wherein the sum of the weights in the long leg is equal to 1 and in the short leg is equal to -1. By applying equation (14), one obtains

$$\sum_{n=1}^N \frac{b_n^{\{LS\}} f_n}{Q_n} = \frac{\sum_{n\in\{L\}} f_n}{\sum_{n\in\{L\}} Q_n} - \frac{\sum_{n\in\{S\}} f_n}{\sum_{n\in\{S\}} Q_n} \neq \frac{q^{\{LS\}}}{Q^{\{LS\}}}. \tag{15}$$

Under the additional assumption that asset-level flows are solely driven by the flow into the long-short portfolio ($f_n = b_n^{\{LS\}} q^{\{LS\}}$ for all n), one has $q^{\{LS\}} = \sum_{n\in\{L\}} f_n = -\sum_{n\in\{S\}} f_n$. However, even under this strong assumption, the last step in (15) is still not an equality, because it is implausible to assume that the total market capitalization of the long-short

portfolio $Q^{\{LS\}}$ satisfies $1/\sum_{n\in\{L\}} Q_n + 1/\sum_{n\in\{S\}} Q_n = 1/Q^{\{LS\}}$.

Due to this error in the denominator, the literature incorrectly interprets the left-hand side of equation (15) as the "percent change in shares outstanding purchased or sold" for long-short portfolios. However, this interpretation is not well-founded, because the underlying calculations in (15) do not support it.

Moreover, I demonstrate a concerning point: the total market capitalization $Q^{\{LS\}}$ of the long-short portfolio is zero. This is because the market portfolio takes out the aggregate supply, so other long-short portfolios are net zero. Therefore, phrases like "buying 1% more of a long-short portfolio" are economically meaningless, as they imply division by zero—a mathematically indefensible operation. To see this, I apply equation (11) to total market capitalization. Let $\mathbf{Q} = (Q_1, Q_2, \ldots, Q_N)^\top$ represent the total market capitalization of N assets. The total market capitalization of the portfolios is

$$\mathbf{Q}^{\{\text{portfolio}\}} = \text{var}(\mathbf{B}^\top\mathbf{R})^{-1}\text{cov}(\mathbf{B}^\top\mathbf{R}, \mathbf{Q}^\top\mathbf{R}) = \text{var}(\mathbf{B}^\top\mathbf{R})^{-1}\text{cov}(\mathbf{B}^\top\mathbf{R}, (\mathbf{b}^{\{\text{MKT}\}})^\top\mathbf{R})\sum_{n=1}^{N} Q_n, \tag{16}$$

where the last step uses the fact that the market portfolio weight $\mathbf{b}^{\{\text{MKT}\}}$ is proportional to the total market capitalization $\mathbf{Q}$. This equation shows that the total market capitalization of any portfolio is the return beta of the market portfolio to that portfolio, multiplied by the total market capitalization of all assets.

For the market portfolio, its return beta to itself is by definition one. So the market portfolio's total market capitalization $Q^{\{\text{MKT}\}}$ equals the sum of the capitalization across all assets $\sum_{n=1}^{N} Q_n$, a result that aligns with common intuitions. However, most long-short portfolios, such as the Fama-French factors, are constructed to achieve zero return correlation with the market. As shown by equation (16), this zero correlation implies that the total market capitalization $Q^{\{LS\}}$ of such long-short portfolios is zero, making phrases like "buying 1% more of a long-short portfolio" economically meaningless.

In summary, by contrasting my formula with the flawed formula in the literature, I highlight the importance of a rigorous theoretical foundation for aggregating flows, particularly for long-short portfolios. Furthermore, since $\Delta Q/Q$ is not economically meaningful for long-short portfolios, the conventional IO approach of defining price multipliers as $(\Delta P/P)/(\Delta Q/Q)$ also fails for long-short portfolios. A major motivation for the remainder of this paper is to address this issue. I achieve this by introducing risk, the central concept of asset pricing, into the characterization of the price/quantity relationship.

5 Factor Model of Price Impacts

This section presents the second step of the model, using factors to characterize the $N \times N$ matrix of self and cross-asset price impacts. Section 5.1 derives the factor model of price impacts. Section 5.2 explains how this model improves upon the literature's quadratic-normal models. Section 5.3 highlights the rationale behind the selection of factors as the unique optimal choice.

5.1 Model Derivation

This section proposes a set of axiomatic assumptions to derive the factor model of the F-SDF and price impacts.

ASSUMPTION 1. ***Linearity of price impact***

For any $a_1 \in \mathbb{R}$, $a_2 \in \mathbb{R}$, $\mathbf{f}_1 \in \mathbb{R}^N$, *and* $\mathbf{f}_2 \in \mathbb{R}^N$, *I assume*

$$\Delta\mathbf{p}(a_1\mathbf{f}_1 + a_2\mathbf{f}_2) = a_1\Delta\mathbf{p}(\mathbf{f}_1) + a_2\Delta\mathbf{p}(\mathbf{f}_2). \tag{17}$$

The linearity of price impact is an assumption, not a consequence of the LOOP. The LOOP does not apply when comparing economies with different flows $\mathbf{f}_1$ and $\mathbf{f}_2$. Linearity implies that price impact doubles if flow doubles and that price impact is additive for two flows. The linearity assumption is substantiated by both theoretical and empirical rationales:

theoretically, it nests quadratic-normal models in the literature; empirically, researchers often approximate price impacts as linear, at least locally, when estimating the price's sensitivity to flow.

ASSUMPTION 2. ***Positive compensation for risk***

For any $\mathbf{f} \neq \mathbf{0}$*, I assume* $\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f}) > 0$.

Recall that equation (10) demonstrates that $\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})$ quantifies the arbitrageur's risk compensation for absorbing flow $\mathbf{f}$. Assumption 2 asserts that this compensation must be strictly positive for any non-zero flow. Intuitively, this implies that when there is an inflow into an asset, the asset's price should increase rather than decrease.

Furthermore, as shown in Hansen and Jagannathan (1991), while multiple SDFs can explain a specific set of prices, only one SDF is uniquely defined within the return space. Analogously, in my theory, while multiple F-SDFs can explain a specific set of price impacts, only one F-SDF is uniquely defined within the demeaned return space spanned by $\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]$. Consequently, the projection onto the demeaned return space is useful to streamline the characterization of the F-SDF. (Section 6.1 provides a deeper characterization of this projection and its connection to the traditional theory.) The aforementioned assumptions allow for a linear expression of the F-SDF. Appendix A.1 contains a proof.

PROPOSITION 2. *There exists a unique F-SDF that satisfies the following conditions:*

i. pricing equation: $\mathbb{E}[\Delta M(\mathbf{f})\mathbf{R}_0] = \Delta\mathbf{p}(\mathbf{f})$ *for every* $\mathbf{f} \in \mathbb{R}^N$*;*

ii. zero-mean property: $\mathbb{E}[\Delta M(\mathbf{f})] = 0$ *for every* $\mathbf{f} \in \mathbb{R}^N$*;*

iii. projection: $\Delta M(\mathbf{f})$ *belongs to the return space spanned by* $\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]$ *for every* $\mathbf{f} \in \mathbb{R}^N$*;*

iv. linearity in Assumption 1.

v. positive compensation for risk in Assumption 2.

Specifically, this unique F-SDF can be written as

$$\Delta M(\mathbf{f}) = (\mathbf{Y}\mathbf{f})^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]), \tag{18}$$

where $\mathbf{Y}$ *is an* $N \times N$ *matrix that ensures* $\text{var}(\mathbf{R}_0)\mathbf{Y}$ *is positive definite.*

The F-SDF $\Delta M(\mathbf{f})$ in (18) has N^2 free parameters $\mathbf{Y}$, and implies a price impact model

$$\Delta \mathbf{p}(\mathbf{f}) = \mathbb{E}[\Delta M(\mathbf{f})\mathbf{R}_0] = \text{var}(\mathbf{R}_0)\mathbf{Y}\mathbf{f}. \tag{19}$$

Intuitively, the model (19) contains the right components, as $\text{var}(\mathbf{R}_0)$ represents fundamental payoff risk and $\mathbf{f}$ corresponds to flow. However, the $N \times N$ matrix $\mathbf{Y}$ as free parameters means that flow into any asset can impact the price of any other asset under this model. This matrix $\mathbf{Y}$ is both hard to interpret theoretically and estimate empirically. Hence, the remainder of this section introduces an economic restriction that uses factors to characterize the $N \times N$ price impacts.

Doing so requires taking a stance on what types of cross-asset price impacts are plausible and what types are not. I examine two mechanisms. The first, supported by empirical findings from Andrade, Chang, and Seasholes (2008) and Chaudhary, Fu, and Li (2023), is correlated payoffs. For instance, should noise traders purchase Google, arbitrageurs selling Google might buy Facebook to hedge against technology sector risk, creating a cross-impact. The second mechanism is correlated flows, backed by studies from Hasbrouck and Seppi (2001) and Ben-David, Li, Rossi, and Song (2022b), where noise traders simultaneously buying or selling assets like Google and Facebook induce cross-impacts.

Conversely, for two assets with both uncorrelated flows $\mathbf{f}$ and uncorrelated fundamental returns $\mathbf{R}_0$, one should not expect the flow into the first asset to impact the price of the second asset. The following Irrelevance of Uncorrelated Flows (IUF) assumption formalizes this intuition. While the economic logic is simple, the technical specifics of IUF are more involved, which I will examine next.

ASSUMPTION 3. ***Irrelevance of Uncorrelated Flows (IUF)***

I define the set of matrices,

$$\mathcal{C} := \left\{ \mathbf{C} \in \mathbb{R}^{N \times N} \middle| \, \text{var}(\mathbf{R}_0)\text{var}(\mathbf{f})\mathbf{C} = \mathbf{C}\text{var}(\mathbf{f})\text{var}(\mathbf{R}_0) \right\}. \tag{20}$$

For any given portfolio $\mathbf{a} \in \mathbb{R}^N$ *and flow* s*, I construct the portfolio,*

$$\mathbf{d} = (\text{cov}(f_1, s), \text{cov}(f_2, s), \ldots, \text{cov}(f_N, s))^\top. \tag{21}$$

If $\mathbf{a}^\top \mathbf{C} \mathbf{d} = 0$ *for all* $\mathbf{C} \in \mathcal{C}$*, then* $\text{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\mathbf{f}), s) = 0$*.*

The IUF posits that there should be no cross-impacts between two portfolios with weights $\mathbf{a} \in \mathbb{R}^N$ and $\mathbf{d} \in \mathbb{R}^N$, if they meet certain conditions. Here, s denotes the flow into portfolio $\mathbf{d}$, and thus $\text{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\mathbf{f}), s) = 0$ implies that the flow into portfolio $\mathbf{d}$ does not affect the price changes in portfolio $\mathbf{a}$.

The complexity of the IUF arises in detailing the conditions that the portfolios $\mathbf{a}$ and $\mathbf{d}$ must satisfy. While the core concept requires these portfolios to exhibit uncorrelated flows and fundamental returns, the scenario is more complex in multi-asset markets due to the possibility of indirect correlations via a third asset. To be precise, portfolios $\mathbf{a}$ and $\mathbf{d}$ are required to be in distinct eigenspaces corresponding to both the fundamental returns $\mathbf{R}_0$ and the flows $\mathbf{f}$.

The application of commuting matrices is pivotal in this analysis. Specifically, square matrices $\mathbf{A}$ and $\mathbf{B}$ commute (i.e., $\mathbf{AB} = \mathbf{BA}$) if and only if they share the same eigenvectors. This property underpins the construction of the matrix set $\mathcal{C}$ as defined in equation (20). Essentially, each matrix $\mathbf{C}$ within $\mathcal{C}$ represents a potential way that portfolios $\mathbf{a}$ and $\mathbf{d}$ could correlate via a third portfolio. The condition "$\mathbf{a}^\top \mathbf{C} \mathbf{d} = 0$ for all $\mathbf{C} \in \mathcal{C}$" ensures that portfolios $\mathbf{a}$ and $\mathbf{d}$ are truly uncorrelated, both in terms of fundamental returns and flows, even when accounting for all possible indirect correlations.

Another detailed aspect of the IUF concerns how portfolio flow s aligns with port-

folio weights $\mathbf{d}$, as shown by equation (21). The sensitivity of asset flow f_n to portfolio flow s is quantified by the beta coefficient $\text{cov}(f_n, s)/\text{var}(s)$. The portfolio weights $\mathbf{d} = (d_1, d_2, \ldots, d_n)^\top$ specify how a one-dollar flow into the portfolio gets allocated: d_n dollars go to asset n. Hence, by setting $d_n = \text{cov}(f_n, s)/\text{var}(s)$, $\mathbf{d}$ becomes the exact portfolio into which s flows. To further simplify equation (21), we normalize all portfolio weights by $\text{var}(s)$. This concept of portfolio weights being the covariance between asset flows and portfolio flows also appears in equation (13), which is part of the broader discussion on portfolio flow theory.

Using IUF to characterize F-SDF and price impacts. Following the explanation of the IUF, I show how it leads to a factor model for the F-SDF and price impacts. To simplify exposition in the main text, I introduce a technical assumption.

ASSUMPTION 4. *Let the Cholesky decomposition of the fundamental risk matrix be* $\text{var}(\mathbf{R}_0) = \mathbf{U}^\top\mathbf{U}$*, where* $\mathbf{U}$ *is an* $N \times N$ *upper triangular matrix. I assume that the matrix* $\text{var}(\mathbf{Uf})$ *has distinct eigenvalues.*

Recall that in introductory linear algebra courses, distinct eigenvalues typically lead to general and straightforward cases. In contrast, duplicate eigenvalues are associated with more complex and nuanced cases. This principle also applies to my theory. Online Appendix A studies scenarios where $\text{var}(\mathbf{Uf})$ may have duplicate eigenvalues. Under such circumstances, the IUF does not restrict how flows impact the price of portfolios within the same eigenspace. This occurs because the commonality in flows and fundamental risks fails to uniquely pin down N asset-pricing factors. Consequently, the theory acknowledges data limitations, and the resulting price impact model is less restrictive.

The following lemma shows that, regardless of the correlation structures in flows and fundamental returns for the N assets, one can always uniquely identify N factor portfolios with uncorrelated flows and uncorrelated fundamental returns. As will be shown, the IUF implies that these specific portfolios have no cross-impacts with each other, so they serve as the basis for constructing the factor model for the F-SDF and price impacts. Appendix A.2

provides a proof.

LEMMA 1. *There exist N uniquely defined portfolios (up to sign) $\mathbf{b}_n = (b_{1,n}, b_{2,n}, \ldots, b_{N,n})^\top$, satisfying:*

i. Factor decomposition of flow:

$$\mathbf{f} = \sum_{n=1}^{N} \mathbf{b}_n q_n, \tag{22}$$

where q_n is the flow into portfolio n.

ii. Uncorrelated fundamental risk:

$$\operatorname{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{b}_m^\top \mathbf{R}_0) = 0 \text{ for all } n \neq m, \text{ and } \operatorname{var}(\mathbf{b}_n^\top \mathbf{R}_0) = 1 \text{ for all } n. \tag{23}$$

iii. Uncorrelated flow:

$$\operatorname{cov}(q_n, q_m) = 0 \text{ for all } n \neq m, \text{ and } \operatorname{var}(q_n) = \pi_n \text{ for all } n. \tag{24}$$

where $\pi_1 > \pi_2 > \cdots > \pi_N > 0$.

I refer to the portfolios $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$ as *factor portfolios* and the corresponding flows $q_1, q_2, \ldots, q_N$ as *factor flows*. The factor decomposition (22) justifies this interpretation, which follows from the portfolio theory introduced in equation (13). As shown in equation (23), factor portfolios have uncorrelated and unit fundamental risk. Furthermore, equation (24) shows that factor flows are uncorrelated with variances ordered from largest to smallest. The uniqueness of these factors arises from their construction, which is explicitly designed to orthogonalize both the flows and fundamental returns.

To better understand these factors, I compare my approach with the standard principal component analysis (PCA) performed on the flows $\mathbf{f} = \sum_{n=1}^{N} \mathbf{b}_n q_n$. Both approaches require uncorrelated flows as in equation (24). However, the standard PCA imposes the orthogonalization condition $\mathbf{b}_n^\top \mathbf{b}_m = 0$ for all $n \neq m$, ensuring that the portfolio weights of different

factors are orthogonal. This condition, unfortunately, does not hold economic significance. In contrast, the factors in this paper are constructed under the condition $\text{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{b}_m^\top \mathbf{R}_0) = 0$ for all $n \neq m$, which ensures that the fundamental returns of different factors are uncorrelated. It is evident that orthogonal portfolio risks are more economically meaningful than orthogonal portfolio weights.

The following theorem shows that if the IUF assumption holds, then the F-SDF can be succinctly expressed using the factors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$. Conversely, if the F-SDF is expressible in this form, then the IUF must hold. Simply put, the IUF completely characterizes the factor model, and vice versa. Appendix A.3 provides a proof.

THEOREM 1. *The F-SDF satisfies all restrictions in Proposition 2 and the IUF Assumption 3* ***if and only if*** *it can be written as*

$$\Delta M(\mathbf{f}) = \sum_{n=1}^{N} \lambda_n q_n \mathbf{b}_n^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) \tag{25}$$

for some parameters $\lambda_n > 0$ for all n. The corresponding price impact model is

$$\Delta \mathbf{p}(\mathbf{f}) = \mathbb{E}[\Delta M(\mathbf{f}) \mathbf{R}_0] = \sum_{n=1}^{N} \lambda_n q_n \text{cov}(\mathbf{R}_0, \mathbf{b}_n^\top \mathbf{R}_0). \tag{26}$$

To understand the factor model of the F-SDF (25) and price impacts (26), it is helpful to first examine the price impacts of the factors $\mathbf{b}_n$. Because different factors have uncorrelated fundamental returns as required by equation (23), I have

$$\mathbf{b}_n^\top \Delta \mathbf{p}(\mathbf{f}) = \lambda_n q_n \text{var}(\mathbf{b}_n^\top \mathbf{R}_0). \tag{27}$$

The key thing to note is that the price impacts of the factor $\mathbf{b}_n$ depend only on its own flow q_n and fundamental return $\mathbf{b}_n^\top \mathbf{R}_0$, but not on other factors. This is the key economic restriction imposed by the IUF assumption—no cross-impacts between the carefully constructed factors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$. This restriction allows reducing the number of free pa-

rameters from N^2 in Proposition 2 to N in Theorem 1. This simplification not only makes the model more tractable but also more robust by reducing the potential for overfitting and enhancing the interpretability of the factors.

It is also important to examine the parameter λ_n. As discussed in Section 4.3, the traditional price multipliers cannot be applied to long-short portfolios. The parameter λ_n in equation (27) is a proper measure for factor-level price response to flows. Specifically, λ_n measures how much the factor price changes for each unit of risk $\text{var}(\mathbf{b}_n^\top \mathbf{R}_0)$ induced by the flow q_n. Hence, λ_n is termed "the price of flow-induced risk." For this parameter, it is crucial to explicitly measure risk exposures, because flows impact prices by changing the risk exposures of arbitrageurs. Recall that traditional asset pricing focuses on the price of risk, expressed as the ratio of expected returns to risk exposures. The price of flow-induced risk λ_n represents the sensitivity of the traditional price of risk to the flow for each factor.

Having established the factor-level price impacts in (27), the derivation of asset-level price impacts as shown in (26) follows from the standard arbitrage-pricing logic. Specifically, individual asset prices adjust in response to changes in factor prices, reflecting their respective risk exposures to the factors, $\text{cov}(\mathbf{R}_0, \mathbf{b}_n^\top \mathbf{R}_0)$.

One can express the F-SDF in (25) and the price impact model (26) using a different set of factors than the ones constructed in Lemma 1. However, because those factors generally do not have uncorrelated flows and uncorrelated fundamental returns, they still have cross-impacts between each other. The IUF implies a structural restriction on these cross-impacts, and the resulting model is more complicated. Appendix B provides the model under alternative factors.

5.2 Improvement Upon Quadratic-Normal Models

In this section, I show how my model improves upon the literature's quadratic-normal models.

While my arbitrage approach to price impacts does not require any parametric assumptions on utility function and the payoff distribution, the literature typically resorts

to quadratic-normal setups. Specifically, the utility function is assumed to be quadratic or CARA, while the payoff distribution is assumed to be multivariate normal. Appendix C shows that the static quadratic-normal price impact model is a special case of my factor model (26), with an additional restriction that $\lambda_1 = \lambda_2 = \cdots = \lambda_N = \gamma/(\mu R_F)$, where γ is the arbitrageurs' risk aversion parameter, μ is the mass of arbitrageurs, and R_F is the gross risk-free rate.

The price of flow-induced risk λ_n endogenously emerges through my arbitrage approach. It also gets at the core of equilibrium models: the ratio of risk aversion to the mass of arbitrageurs. The static quadratic-normal model requires this ratio to be the same for all factors. Not only is this restriction empirically rejected by An, Su, and Wang (2023), but it also disappears in a dynamic quadratic-normal model with predictable flows, as I show in Section 7.1. My factor model (26) improves upon the quadratic-normal model by allowing heterogeneous price of flow-induced risk among different factors. Theorem 1 shows that, even without the uniform price of flow-induced risk, the quadratic-normal model retains a structured $N \times N$ price impact matrix. The IUF condition precisely characterizes this structure.

Simply put, the static quadratic-normal model uses one free parameter to characterize the $N \times N$ cross-substitution patterns, while my model (26) uses N parameters. In the static quadratic-normal model, the cross-substitution between two assets depends on their factor risk exposures, but these factors must have the same price of flow-induced risk. In my model, the factors' price of flow-induced risk can differ from each other. Therefore, some cross-asset price impacts that do not show up in the static quadratic-normal model emerge under my model. The intuition is that flow into asset A can impact the price of asset B by differentially impacting different factors that asset B is exposed to.

I now present the precise mathematics for this new channel of cross-impacts. I consider N portfolios $\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N$ with uncorrelated fundamental returns, i.e., $\operatorname{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ for $n \neq m$. Unlike the factors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$ constructed in Theorem 1, the portfolios

$\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N$ may have correlated flows. The question is, holding all else equal, does flow $\tilde{q}_m$ into portfolio m impact the price $\tilde{\mathbf{b}}_n^\top \Delta\mathbf{p}(\mathbf{f})$ of portfolio n for $m \neq n$? Importantly, "holding all else equal" controls for the portfolio-n flow $\tilde{q}_n$, and thus eliminates the mechanical cross-impact arising from $\tilde{q}_m$ changing $\tilde{q}_n$. My model answers yes to this question, while the quadratic-normal model answers no.

To see this, by Proposition 5, the price impact of portfolio $\tilde{\mathbf{b}}_n$ under my factor model is

$$\tilde{\mathbf{b}}_n^\top \Delta\mathbf{p}(\mathbf{f}) = \sum_{m=1}^{N} \tilde{\lambda}_{n,m} \tilde{q}_m \mathrm{var}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0), \tag{28}$$

where the price-of-flow-induced-risk matrix $\tilde{\mathbf{\Lambda}} = \{\tilde{\lambda}_{n,m}\}$ is given by

$$\tilde{\mathbf{\Lambda}} = \mathbf{O}\mathbf{\Lambda}\mathbf{O}^{-1} = \mathbf{O}\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)\mathbf{O}^{-1}. \tag{29}$$

The matrix $\mathbf{O}$ rotates the portfolios $\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N$ to the factor portfolios $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$ that have both uncorrelated fundamental returns and uncorrelated flows in Theorem 1. Equation (29) shows that the off-diagonal term $\tilde{\lambda}_{n,m}$ is generally nonzero for $m \neq n$. Therefore, by (28), flow $\tilde{q}_m$ into portfolio m impacts the price $\tilde{\mathbf{b}}_n^\top \Delta\mathbf{p}(\mathbf{f})$ of portfolio n.

The static quadratic-normal model, requiring $\lambda_1 = \lambda_2 = \cdots = \lambda_N$, is a special case of my factor model. In this case, equation (29) implies that

$$\tilde{\mathbf{\Lambda}} = \mathbf{O}\mathrm{diag}(\lambda_1, \lambda_1, \ldots, \lambda_1)\mathbf{O}^{-1} = \mathrm{diag}(\lambda_1, \lambda_1, \ldots, \lambda_1) \tag{30}$$

is diagonal, i.e., $\tilde{\lambda}_{n,m} = 0$ for $m \neq n$. Therefore, flow $\tilde{q}_m$ into portfolio m does not impact the price $\tilde{\mathbf{b}}_n^\top \Delta\mathbf{p}(\mathbf{f})$ of portfolio n. The absence of cross-impacts is because flow $\tilde{q}_m$ impacts different factors that portfolio n is exposed to in the same magnitude λ_1.

Comparing equations (29) and (30), one sees that whether the cross-asset price impacts arise depends precisely on whether all λ_n are equal to each other, which is the key improvement of my factor model over the quadratic-normal model. Hence, by empirically rejecting

the hypothesis that all λ_n are equal, An, Su, and Wang (2023) support the new channel.

In summary, while the static quadratic-normal model rules out cross-impacts between any two portfolios with uncorrelated fundamental returns, even in cases where these portfolios have correlated flows, my IUF approach adopts a more general stance. It only eliminates cross-impacts between portfolios that exhibit both uncorrelated fundamental returns and uncorrelated flows, thereby enriching the cross-substitution pattern. In a similar vein, Buffa and Hodor (2023) show that cross-impacts can occur between assets with uncorrelated fundamental returns in scenarios featuring heterogeneous benchmarking. Furthermore, the cross-substitution pattern implied by my IUF model markedly differs from that of the logit demand system in Koijen and Yogo (2019). For instance, according to my IUF model, a flow into an asset A with a positive beta would decrease the price of another asset B with a negative beta, reflecting the arbitrageurs' risk hedging incentive. Conversely, the logit demand system, which does not account for return covariances, would suggest an increase in the price of asset B due to its proportional cross-substitution pattern.[19]

5.3 Optimality of Chosen Factors

Having established that the IUF assumes no cross-impacts among certain factors, it stands out for selecting factors with uncorrelated flows and uncorrelated fundamental returns. While this selection is empirically grounded in two cross-impact channels, this section presents a deeper theoretical justification that stands independently of any specific channel. I prove that, out of all possible ways of selecting N factors and eliminating cross-impacts among them, the IUF is the only way that keeps each factor's price response to its flow unchanged and keeps the arbitrageur's expected utility unchanged for any specific utility function. This section is developed under the regularity Assumption 4, while Online Appendix A presents the general theory without this assumption.

[19]Proportional substitution is a well-known issue of the logit demand system. For a detailed discussion, refer to Section 3.3.2 of Train (2009).

Revisiting the N^2 model (19), I first define the process of selecting N factors and eliminating cross-impacts among them, termed "model orthogonalization."

DEFINITION 1. *Any N linearly independent portfolios $\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N$ defines a model orthogonalization. The N^2 model (19) expressed under these portfolios is*

$$\Delta \mathbf{p}(\mathbf{f}) = \sum_{n=1}^{N} \sum_{m=1}^{N} \tilde{\lambda}_{n,m} \tilde{q}_m \mathrm{cov}(\mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0), \tag{31}$$

with the N^2 free parameters $\tilde{\lambda}_{n,m}$ and portfolio flows $\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_N$. The orthogonalized N-factor model under these portfolios is defined as

$$\Delta \bar{\mathbf{p}}(\mathbf{f}) = \sum_{n=1}^{N} \tilde{\lambda}_{n,n} \tilde{q}_n \mathrm{cov}(\mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0). \tag{32}$$

By Definition 1, model orthogonalization means that one picks a specific set of N portfolios $\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N$ and removes the off-diagonal terms of the corresponding $\tilde{\lambda}_{n,m}$. By Theorem 1, the IUF is a particular model orthogonalization that chooses the specific factor portfolios $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$ with uncorrelated fundamental returns and uncorrelated flows.

To show why the IUF is the optimal orthogonalization, I next define the expected utility of an arbitrageur. Denote the arbitrageur's wealth after absorbing flow $\mathbf{f}$ as $W(\mathbf{f})$, and the arbitrageur's utility function on her wealth as $u(\cdot)$. By absorbing flow $\mathbf{f}$, the arbitrageur's expected utility is given by

$$\mathbb{E}[u(W(\mathbf{f}))] = \mathbb{E}[u(W(\mathbf{0}) + R_F \mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f}) - \mathbf{f}^\top (\mathbf{R}_0 - R_F \mathbf{1}))], \tag{33}$$

where I use equation (10). As discussed in Section 4.1, $\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f})$ is the arbitrageur's compensation for absorbing the flow-induced risk, and $\mathbf{f}^\top (\mathbf{R}_0 - R_F \mathbf{1})$ is the total payoff risk.

Because flow $\mathbf{f}$ can vary, the compensation for risk $\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f})$ also varies. However, because flow $\mathbf{f}$ is independent of the assets' payoff $\mathbf{X}$, the compensation $\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f})$ is also independent of $\mathbf{X}$. Hence, to eliminate the second-order effect of the arbitrageur's aversion to the varying

compensation $\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})$, I assume that the arbitrageur is averse only to the fundamental payoff risk, i.e., for some utility function $\tilde{u}(\cdot)$,

$$u(W(\mathbf{0}) + R_F \mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f}) - \mathbf{f}^\top(\mathbf{R}_0 - R_F\mathbf{1})) = R_F \mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f}) + \tilde{u}(W(\mathbf{0}) - \mathbf{f}^\top(\mathbf{R}_0 - R_F\mathbf{1})). \quad (34)$$

Using equations (33) and (34), I simplify the arbitrageur's expected utility as

$$\mathbb{E}[u(W(\mathbf{f}))] = R_F \mathbb{E}[\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})] + \mathbb{E}[\tilde{u}(W(\mathbf{0}) - \mathbf{f}^\top(\mathbf{R}_0 - R_F\mathbf{1}))]. \quad (35)$$

The insight from (35) is that the price impact $\Delta\mathbf{p}(\mathbf{f})$ affects the expected utility $\mathbb{E}[u(W(\mathbf{f}))]$ only by varying the expected compensation $\mathbb{E}[\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})]$. Therefore, by holding $\mathbb{E}[\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})]$ constant while reducing the N^2 model to an N-factor model, I can ensure that the expected utility $\mathbb{E}[u(W(\mathbf{f}))]$ remains invariant for any utility function. Theorem 2 applies this insight and shows that the IUF indeed keeps $\mathbb{E}[\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})]$ invariant. Appendix A.4 provides a proof.

THEOREM 2. *Fix any model orthogonalization $\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N$. The orthogonalized model (32) satisfies the IUF for any N parameters $\tilde{\lambda}_{1,1}, \tilde{\lambda}_{2,2}, \ldots, \tilde{\lambda}_{N,N}$* ***if and only if*** *for any N^2 parameters $\tilde{\lambda}_{n,m}$,*

- *a one-unit shock to portfolio flow $\tilde{q}_n$ causes the same amount of impact to the price of portfolio $\tilde{\mathbf{b}}_n$ under the N^2 model (31) and the orthogonalized model (32),*

$$\frac{\partial \tilde{\mathbf{b}}_n^\top \Delta\mathbf{p}(\mathbf{f})}{\partial \tilde{q}_n} = \frac{\partial \tilde{\mathbf{b}}_n^\top \Delta\bar{\mathbf{p}}(\mathbf{f})}{\partial \tilde{q}_n}. \quad (36)$$

- *the arbitrageur's expected utility $\mathbb{E}[u(W(\mathbf{f}))]$ remains invariant under (31) and (32).*

Similar to Theorem 1, Theorem 2 is also an if-and-only-if statement, so it provides an alternative characterization of the IUF based on the optimality of chosen factors. That is, the IUF is the only model orthogonalization that satisfies two key properties. First, (36) ensures that each factor's price response to its own flow remains invariant for the N^2 model

and the orthogonalized N-factor model. Empirically, this property ensures that the price impact regression for the orthogonalized model correctly recovers each factor's $\tilde{\lambda}_{n,n}$.

Second, model orthogonalization eliminates off-diagonal terms of $\tilde{\lambda}_{n,m}$, and as a result, it necessarily loses some information. The IUF is the unique optimal orthogonalization in the sense that it preserves the arbitrageur's expected utility $\mathbb{E}[u(W(\mathbf{f}))]$, which is the arbitrageur's optimizing goal in equilibrium models. The theorem's theoretical generality lies in the fact that it imposes no parametric assumptions on the arbitrageur's utility function, yet demonstrates that the IUF maintains the invariance of expected utility across both the N^2 model and the orthogonalized N-factor model.

6 Reducing the Number of Factors

This section presents the last step of the model, where I reduce the number of factors from N to a small K in equation (26). The motivation is that in practice, N is usually large, and empirical researchers often use a small set of factors to achieve robustness in estimation. To do so, Section 6.1 generalizes the concept of arbitrage to markets with noise trading. Section 6.2 applies this concept and the fact that flows themselves exhibit a low-dimensional factor structure to reduce the number of factors.

6.1 Generalized Concept of Arbitrage

Traditionally, no arbitrage means that a portfolio with little risk should not have a high expected return. The Hansen-Jagannathan bound acts as a bridge, linking the maximum attainable Sharpe ratio within the economy to the volatility of the SDF. I generalize this principle for noise trading scenarios, postulating that a small flow into a portfolio with little payoff risk should not generate a high price impact. If such portfolios existed, it means that only a limited number of arbitrageurs are providing liquidity, granting them exceptionally favorable risk-return trade-offs.

This section delves into this generalized concept of arbitrage, developing the flow-based counterparts to both the Sharpe ratio and the Hansen and Jagannathan (1991) bound. Note that this generalized arbitrage extends beyond the scope of the previously introduced factor model. Specifically, the results derived in this section require only the assumptions made on model setup in Section 3.

By the pricing equation (8), the price impact of any portfolio $\mathbf{c} \in \mathbb{R}^N$ is $\mathbf{c}^\top \Delta \mathbf{p}(\mathbf{f}) = \mathbb{E}[\mathbf{c}^\top \mathbf{R}_0 \Delta M(\mathbf{f})]$. Using the F-SDF's zero-mean property from (9), I have

$$\mathbb{E}[\mathbf{c}^\top \mathbf{R}_0 \Delta M(\mathbf{f})] = \text{corr}(\mathbf{c}^\top \mathbf{R}_0, \Delta M(\mathbf{f}))\sigma(\mathbf{c}^\top \mathbf{R}_0)\sigma(\Delta M(\mathbf{f})), \tag{37}$$

where $\sigma(\cdot)$ denotes volatility. Because the correlation $|\text{corr}(\mathbf{c}^\top \mathbf{R}_0, \Delta M(\mathbf{f}))| \leq 1$ for any portfolio $\mathbf{c}$, I have

$$\sigma(\Delta M(\mathbf{f})) \geq \max_{\mathbf{c} \in \mathbb{R}^N} \frac{\mathbf{c}^\top \Delta \mathbf{p}(\mathbf{f})}{\sigma(\mathbf{c}^\top \mathbf{R}_0)}. \tag{38}$$

On the right-hand side, the numerator $\mathbf{c}^\top \Delta \mathbf{p}(\mathbf{f})$ represents the portfolio's price impact, while the denominator $\sigma(\mathbf{c}^\top \mathbf{R}_0)$ denotes the portfolio's fundamental-return volatility. Consequently, $\mathbf{c}^\top \Delta \mathbf{p}(\mathbf{f})/\sigma(\mathbf{c}^\top \mathbf{R}_0)$ is termed the *price impact ratio* of portfolio $\mathbf{c}$ under flow $\mathbf{f}$, and the right-hand side of (38) is the *maximum price impact ratio (MPIR)* across all portfolios.

Inequality (38) shows that, for any fixed flow $\mathbf{f}$, the F-SDF's volatility is no less than the maximum price impact ratio. Appendix A.5 proves that this bound is tight.

PROPOSITION 3. *The flow-based Hansen-Jagannathan bound is*

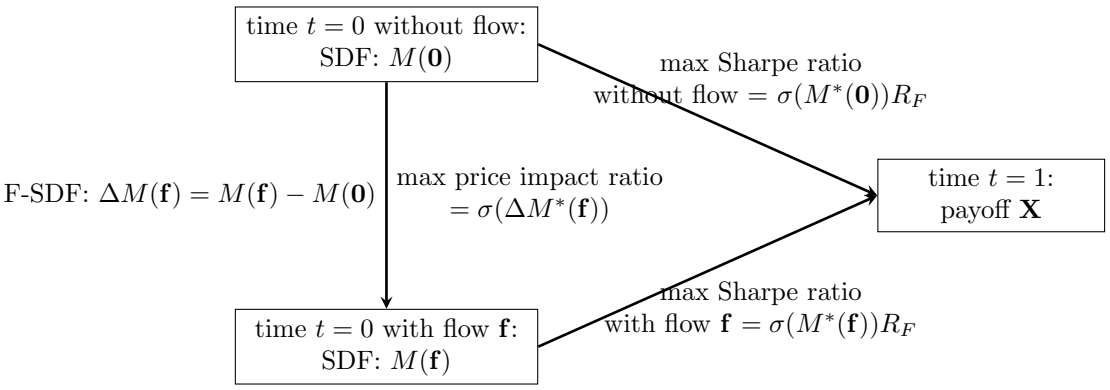
$$\min_{\Delta M(\mathbf{f})} \sigma(\Delta M(\mathbf{f})) = \max_{\mathbf{c} \in \mathbb{R}^N} \frac{\mathbf{c}^\top \Delta \mathbf{p}(\mathbf{f})}{\sigma(\mathbf{c}^\top \mathbf{R}_0)} \tag{39}$$

for any given flow $\mathbf{f}$*. Specifically, the minimum-volatility F-SDF is*

$$\Delta M^*(\mathbf{f}) = \Delta \mathbf{p}(\mathbf{f})^\top \text{var}(\mathbf{R}_0)^{-1}(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]), \tag{40}$$

which is also the unique F-SDF in the demeaned return space spanned by $\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]$*.*

Figure 3. Relationship with standard theory



Notes: The F-SDF $\Delta M(\mathbf{f})$ is the difference between two standard SDFs, $M(\mathbf{f})$ and $M(\mathbf{0})$. The original Hansen-Jagannathan bound shows that the volatility of the minimum-volatility SDF $M^*(\mathbf{f})$ and $M^*(\mathbf{0})$ (times the risk-free rate R_F) is equal to the maximum Sharpe ratio in the economy with and without flow $\mathbf{f}$, respectively. The flow-based Hansen-Jagannathan bound shows that the volatility of the minimum-volatility F-SDF $\Delta M^*(\mathbf{f}) = M^*(\mathbf{f}) - M^*(\mathbf{0})$ is equal to the maximum price impact ratio under flow $\mathbf{f}$.

The key of Proposition 3 lies in the portfolio that has the maximum price impact for a given level of fundamental risk. Intuitively, shorting this MPIR portfolio is the mean-variance optimal strategy to capitalize on flow-induced changes in risk premiums. As discussed in Proposition 2, while multiple F-SDFs can explain a specific set of price impacts, only one F-SDF is uniquely defined within the demeaned return space spanned by $\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]$. Proposition 3 further shows that this unique F-SDF is also the one with the minimum volatility, and its volatility actually equals the MPIR of the mean-variance optimal strategy.

Figure 3 illustrates the relationship of my theory with standard theory. Instead of studying the Sharpe ratio per se, my theory focuses on how flows change the Sharpe ratio, which intuitively captures the arbitrageurs' risk-absorbing capacity on the margin. To see this, note that by equations (3) and (4), the minimum-volatility standard SDF under flow $\mathbf{f}$ is[20]

$$M^*(\mathbf{f}) = \frac{1}{R_F} + \left(\Delta\mathbf{p}(\mathbf{f}) - \frac{\mathbb{E}[\mathbf{R}_0 - R_F\mathbf{1}]}{R_F}\right)^\top \mathrm{var}(\mathbf{R}_0)^{-1}(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]). \tag{41}$$

The original Hansen-Jagannathan bound shows that the volatility of $M^*(\mathbf{f})$ and $M^*(\mathbf{0})$ (times the risk-free rate R_F) is equal to the maximum Sharpe ratio in the economy with and

[20]Refer to equation (5.25) in Cochrane (2009).

without flow $\mathbf{f}$, respectively. In my theory, the minimum-volatility F-SDF $\Delta M^*(\mathbf{f})$ equals the the difference between $M^*(\mathbf{f})$ and $M^*(\mathbf{0})$. The flow-based Hansen-Jagannathan bound shows that the volatility of this F-SDF $\Delta M^*(\mathbf{f})$ is equal to the MPIR.

6.2 Flow-Based APT

Equipped with the generalized concept of arbitrage, this section develops the flow-based APT to reduce the N-factor model of price impacts (26) to a low-dimensional K-factor model.

I write the N-asset flow as the N-factor structure $\mathbf{f} = \sum_{n=1}^{N} \mathbf{b}_n q_n$ as in Lemma 1, with factor portfolios $\mathbf{b}_n$ and corresponding flows q_n sorted such that their variance satisfies $\pi_1 > \pi_2 > \cdots > \pi_N$. The N-factor price impact model (26) is $\Delta\mathbf{p}(\mathbf{f}) = \sum_{n=1}^{N} \lambda_n q_n \text{cov}(\mathbf{R}_0, \mathbf{b}_n^\top \mathbf{R}_0)$. The flow-based APT uses the first K factors, $\mathbf{f} = \sum_{k=1}^{K} \mathbf{b}_k q_k + \mathbf{e}$, where $\mathbf{e}$ is an $N \times 1$ vector of idiosyncratic flows. The reason to pick first K factors is because empirically, flows indeed exhibit a low-dimensional factor structure (Hasbrouck and Seppi, 2001). The corresponding K-factor price impact model is

$$\Delta\check{\mathbf{p}}(\mathbf{f}) = \sum_{k=1}^{K} \lambda_k q_k \text{cov}(\mathbf{R}_0, \mathbf{b}_k^\top \mathbf{R}_0). \tag{42}$$

Under what conditions can I bound the pricing error $\Delta\mathbf{p}(\mathbf{f}) - \Delta\check{\mathbf{p}}(\mathbf{f})$ of price impacts? The following proposition provides the answer, and Appendix A.6 presents a proof.

PROPOSITION 4. *Let $\|\mathbf{v}\|$ be the L^2 norm of vector $\mathbf{v}$. For any F-SDF that satisfies*

$$\max_{\|\mathbf{f}\|=1} \sigma(\Delta M(\mathbf{f})) \leq H \tag{43}$$

for some constant H, if $\sum_{n=K+1}^{N} \mathbb{E}[q_n^2]$ tends to zero, then $\mathbb{E}[\|\Delta\mathbf{p}(\mathbf{f}) - \Delta\check{\mathbf{p}}(\mathbf{f})\|^2]$ tends to zero.

Condition (43) requires the factor model's F-SDF $\Delta M(\mathbf{f})$ in Theorem 1 not to be overly volatile. Because the F-SDF $\Delta M(\mathbf{f})$ is linear in flow $\mathbf{f}$, I bound the volatility of the F-SDF

on the unit sphere $\|\mathbf{f}\| = 1$. By Proposition 3, the bound on the volatility of the F-SDF is also the bound on the maximum price impact ratio. If the price impact ratio is too high, small flows into some portfolio with little fundamental risks would have a large price impact. These opportunities represent very good deals for arbitrageurs—they can trade against flows to take advantage of price dislocations while taking on little fundamental risks. Condition (43) assumes that these good deals should not exist. This assumption generalizes the "good-deal bound" logic of Cochrane and Saa-Requejo (2000) to markets with noise trading.

Idiosyncratic flows' $\sum_{n=K+1}^{N} \mathbb{E}[q_n^2]$ tending to zero means that flows have a factor structure. Data support this assumption (Hasbrouck and Seppi, 2001). Here, because flows can have nonzero means, $\sum_{n=K+1}^{N} \mathbb{E}[q_n^2]$ tending to zero means that both the variance $\sum_{n=K+1}^{N} \pi_n$ and squared expected flows $\sum_{n=K+1}^{N} \mathbb{E}[q_n]^2$ tend to zero. The term $\Delta\mathbf{p}(\mathbf{f}) - \Delta\check{\mathbf{p}}(\mathbf{f})$ is the pricing error of the price impacts between the true N-factor model and the approximate K-factor model under a given flow $\mathbf{f}$. Because $\mathbf{f}$ is random, $\mathbb{E}[\|\Delta\mathbf{p}(\mathbf{f}) - \Delta\check{\mathbf{p}}(\mathbf{f})\|^2]$ tending to zero implies a good approximation in the mean-squared-error sense. In summary, assuming the F-SDF has bounded volatility, the factor structure of flows implies the factor model of price impacts.

One may question whether a version of the flow-based APT exists that bypasses the IUF and directly simplifies the N^2 model (19) into a K^2 model. Online Appendix B demonstrates that this is not feasible. Without the IUF, there is no solid theoretical basis for disregarding the cross-impacts of factor flows on portfolios receiving idiosyncratic flows. In simpler terms, the K factors identified by the IUF are crucial for the validity of the flow-based APT. These factors are vital not only for capturing commonality in flows but, importantly, for capturing commonality in flow-induced risk exposures.

7 Applications

In this section, I discuss two applications from subsequent works where flows either exhibit dynamic predictability or are informed about asset payoffs. Provided a single-asset model is

solvable, researchers can apply my machinery to solve the general model with N assets featuring any correlation pattern in both payoffs and flows. Unlike traditional approaches that impose stringent correlation structures for tractability, this generality facilitates theoretical investigations into how trading flows impact cross-sectional asset prices in a more realistic setting and delivers new insights.

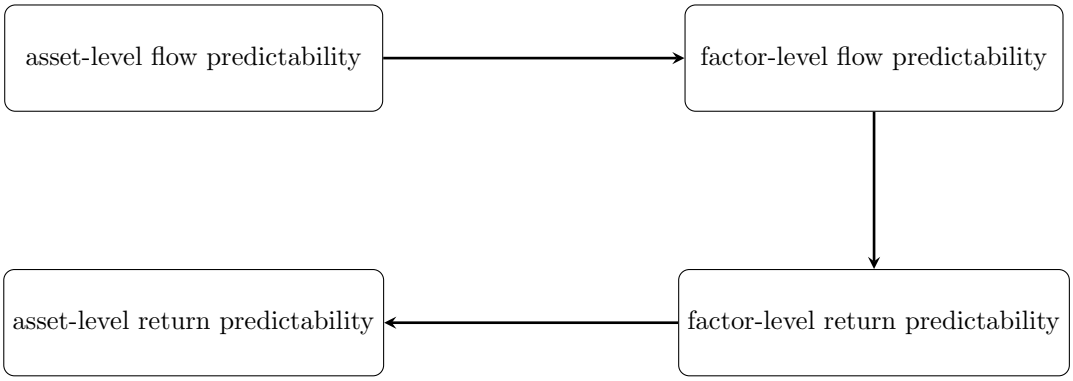
7.1 Predictable Flows

This application examines the cross-sectional asset pricing implications of predicable flows, and also provides an alternative foundation for the factor model of price impacts (26) in a dynamic equilibrium setting. The discussion here largely follows An and Zheng (2023b).

Empirical studies have consistently highlighted the predictability of noise trading flows. A commonly observed pattern is that if noise traders purchase a specific asset today, they are likely to continue purchasing more of the same asset the following day, demonstrating a momentum in flows. This observation leads to an intuitive hypothesis: the predictability of flows, coupled with their influence on asset prices, could partially account for the predictability in returns. This hypothesis receives empirical support from Coval and Stafford (2007) and Lou (2012). Moreover, the relationship between flow predictability and return predictability is not just confined to individual assets but also extends to factors, as documented by Kelly, Moskowitz, and Pruitt (2021), Ben-David, Li, Rossi, and Song (2022a), Ehsani and Linnainmaa (2022), and Arnott, Kalesnik, and Linnainmaa (2023).

These empirical findings prompt a crucial theoretical question: how is flow predictability linked to return predictability at both individual asset and factor levels? Addressing this question requires a dynamic asset pricing model that incorporates three key elements: the covariance structure of fundamental returns, denoted as $\text{var}(\mathbf{R}_0)$; the covariance of noisy flows, $\text{var}(\mathbf{f})$; and the predictability of flows, represented by $\boldsymbol{\kappa}$. In practice, these elements are represented by general $N \times N$ matrices, acknowledging the realities of covariances and cross-predictability observed in the data. Prior studies often simplify their models by assuming at

Figure 4. The dynamic factor model of price impacts



least one of these matrices to be diagonal, effectively eliminating cross-asset relationships.[21]

Building upon the framework in this paper, An and Zheng (2023b) show that in the cross-section, flow predictability drives return predictability through factors. A pivotal aspect of their approach, echoing the significance of commuting matrices in defining the IUF condition in my work, is the strategic use of these matrices in solving the dynamic problem. They show that the matrices $\text{var}(\mathbf{f})$, $\text{var}(\mathbf{R}_0)$, and $\boldsymbol{\kappa}$, exhibit a commuting property (specifically, $\text{var}(\mathbf{f})\text{var}(\mathbf{R}_0)\boldsymbol{\kappa} = \boldsymbol{\kappa}\text{var}(\mathbf{f})\text{var}(\mathbf{R}_0)$) if and only if there is no cross-predictability among the N specific factors identified in Lemma 1. In other words, these N factors have uncorrelated fundamental returns and flows, with each factor's flow predictability acting independently. When flows at the asset level and their predictability are aggregated to these factors, their price impacts can be independently analyzed using the single-asset models of Campbell and Kyle (1993) and Wang (1993). This factor-level flow predictability gives rise to factor-level return predictability, which subsequently cascades down to asset-level return predictability in line with individual assets' risk exposures. This dynamic model is depicted in Figure 4.

As highlighted in Section 5.2, in the static quadratic-normal model, the price of flow-induced risk λ_n is equalized across factors. The dynamic model in An and Zheng (2023b), while adhering to the quadratic-normal framework, relaxes the equal-λ_n constraint through an equilibrium approach. The reason is that flow predictability can differ across factors, and

[21] For an overview of such simplifications, see the survey article by Rostek and Yoon (2020) or the models presented in Bogousslavsky (2016) and Lu, Malliaris, and Qin (2023).

factors with more pronounced flow momentum have higher values of λ_n. The intuition is that stronger flow momentum leads arbitrageurs who absorb such flows today to anticipate more flows tomorrow. This dynamic consideration effectively raises the arbitrageurs' risk aversion for absorbing flows today, leading to a higher λ_n.

In summary, the key advancement of the factor model (26) over the static quadratic-normal model is relaxing the equal-λ_n constraint. My paper highlights this advancement through the arbitrage approach, particularly emphasizing the IUF condition. In a complementary manner, An and Zheng (2023b) explore this advancement from an equilibrium perspective, focusing on the dynamic predictability of flows—a strong pattern observed in the data. This aspect, wherein flow predictability leads to differential λ_n across factors, draws a parallel to the Merton (1973a) ICAPM, which shows that return predictability can introduce additional risk factors beyond the static CAPM.

7.2 Informed Flows

This application examines the cross-sectional asset pricing implications of informed flows, while also offering a broader perspective on the factor model of price impacts (26) in the context of informed trading. The discussion here largely follows An and Zheng (2023a).

Returning to the fundamental pricing equation $P = \mathbb{E}[MX]$, my paper operates under the assumption that the noise trading flow is independent of the payoff X. Consequently, all price impacts are manifested through changes in the SDF M. An and Zheng (2023a) expands upon this framework to accommodate scenarios where the order flow can be informed about the payoff X. This allows arbitrageurs to learn about the payoff from the order flow and determine asset prices P accordingly. Specifically, An and Zheng (2023a) extend the arbitrage-pricing approach developed in my paper to model information learning, diverging from the traditional quadratic-normal learning models in Kyle (1985). This arbitrage approach reveals general constraints on arbitrageurs' learning that are applicable across various reasonable equilibria, and offers three advantages over traditional microstructure approaches:

1) it avoids the limitations of examining learning within specific equilibrium contexts; 2) it removes the need for parametric assumptions about utility functions and payoff distributions; 3) it is versatile enough to accommodate both Bayesian and behavioral learning.

Furthermore, the factors identified by the IUF address a challenge in the literature concerning information learning involving multiple assets with correlated payoffs. The textbook approach (Veldkamp, 2011) is to construct portfolios with uncorrelated payoffs, and assumes that certain agents receive independent signals about the payoff of each portfolio. These informed agents subsequently generate trading flows, from which arbitrageurs learn.

However, the textbook approach has several drawbacks. From an economic standpoint, Sims (2006) critiques this approach, arguing that agents might strategically opt to acquire correlated signals on uncorrelated payoffs. On a technical level, the characterization of such portfolios is not unique—given any set of portfolios with uncorrelated payoffs, it is always possible to rotate them into another set that also have uncorrelated payoffs. This lack of uniqueness implies that the choice of portfolios in this approach depends on an arbitrary orthogonalization condition imposed by theorists.[22] As Veldkamp (2011) notes, "Which procedure most resembles how economic decision makers learn is an open question."

The core concept of this paper, the IUF, directly addresses this open question. As demonstrated in Lemma 1, it is possible to identify unique portfolios characterized by both uncorrelated flows and uncorrelated payoffs, thereby resolving the technical issue. When the IUF is applied to such portfolios, it implies that order flows into one portfolio do not convey information about the payoffs of other portfolios. This is underpinned by the logic of revealed preference. Should agents observe correlated signals on these portfolios, it would lead to correlated order flows, which would contradict the very basis of the portfolio construction. Thus, the IUF provides not only a technical solution but also aligns with economic reasoning, ensuring that the construction of portfolios adheres to logical market behavior.

In summary, while my paper exclusively focuses on the price impacts of noise trading

[22]PCA orthogonalization is commonly used, which lacks a clear economic justification in this context.

flows, the methodologies developed herein are also applicable to the analysis of learning from informed flows, as explored in An and Zheng (2023a). In the expanded model, which still adheres to the factor structure (26), flow can impact price at the factor level through two distinct mechanisms: firstly, by exposing arbitrageurs to risk, and secondly, by conveying information about payoff.

8 Conclusion

In conclusion, this paper proposes an arbitrage-pricing approach to analyze how noise trading flows impact asset prices. This approach uses no-arbitrage conditions to determine the impact of these flows on the stochastic discount factor, and consequently, on the cross-section of asset prices. By generalizing classic arbitrage theory, I show that noisy flows impact asset prices through a few important factors. The resulting model features rich patterns of cross-asset substitution beyond traditional mean-variance price impact models and logit demand systems. My theory also addresses prevalent misconceptions regarding the aggregation of asset flows into factor flows and the definition of factor-level price multipliers.

References

Alvarez, Fernando, and Andrew Atkeson, 2018, The risk of becoming risk averse: A model of asset pricing and trade volumes, Working paper, University of Chicago.

An, Yu, Yinan Su, and Chen Wang, 2023, A factor framework for cross-sectional price impacts, Working paper, Johns Hopkins University.

An, Yu, and Zeyu Zheng, 2023a, An axiomatic approach to informed order flow, Working paper, Johns Hopkins University.

An, Yu, and Zeyu Zheng, 2023b, A dynamic factor model of price impacts, Working paper, Johns Hopkins University.

Andrade, Sandro C, Charles Chang, and Mark S Seasholes, 2008, Trading imbalances, predictable reversals, and cross-stock price pressure, *Journal of Financial Economics* 88, 406–423.

Arnott, Robert D, Vitali Kalesnik, and Juhani T Linnainmaa, 2023, Factor momentum, *Review of Financial Studies* 36, 3034–3070.

Barber, Brad M, Xing Huang, Terrance Odean, and Christopher Schwarz, 2022, Attention-induced trading and returns: Evidence from Robinhood users, *Journal of Finance* 77, 3141–3190.

Ben-David, Itzhak, Jiacui Li, Andrea Rossi, and Yang Song, 2022a, Discontinued positive feedback trading and the decline of return predictability, *Journal of Financial and Quantitative Analysis* Forthcoming.

Ben-David, Itzhak, Jiacui Li, Andrea Rossi, and Yang Song, 2022b, Ratings-driven demand and systematic price fluctuations, *Review of Financial Studies* 35, 2790–2838.

Black, Fischer, and Myron Scholes, 1973, The pricing of options and corporate liabilities, *Journal of Political Economy* 81, 637–654.

Bogousslavsky, Vincent, 2016, Infrequent rebalancing, return autocorrelation, and seasonality, *Journal of Finance* 71, 2967–3006.

Boulatov, Alex, Terrence Hendershott, and Dmitry Livdan, 2013, Informed trading and portfolio returns, *Review of Economic Studies* 80, 35–72.

Buffa, Andrea M, and Idan Hodor, 2023, Institutional investors, heterogeneous benchmarks and the comovement of asset prices, *Journal of Financial Economics* 147, 352–381.

Caballe, Jordi, and Murugappa Krishnan, 1994, Imperfect competition in a multi-security market with risk neutrality, *Econometrica* 695–704.

Campbell, John Y, and Albert S Kyle, 1993, Smart money, noise trading and stock price behaviour, *Review of Economic Studies* 60, 1–34.

Chang, Yen-Cheng, Harrison Hong, and Inessa Liskovich, 2015, Regression discontinuity and the price effects of stock market indexing, *Review of Financial Studies* 28, 212–246.

Chaudhary, Manav, Zhiyu Fu, and Jian Li, 2023, Corporate bond multipliers: Substitutes matter, Working paper, Columbia Business School.

Chordia, Tarun, Richard Roll, and Avanidhar Subrahmanyam, 2000, Commonality in liquidity, *Journal of Financial Economics* 56, 3–28.

Cochrane, John H, 2009, *Asset pricing: Revised edition* (Princeton university press).

Cochrane, John H, and Jesus Saa-Requejo, 2000, Beyond arbitrage: Good-deal asset price bounds in incomplete markets, *Journal of Political Economy* 108, 79–119.

Coval, Joshua, and Erik Stafford, 2007, Asset fire sales (and purchases) in equity markets, *Journal of Financial Economics* 86, 479–512.

De Long, J. Bradford, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann, 1990, Noise trader risk in financial markets, *Journal of Political Economy* 98, 703–738.

Dou, Winston, Leonid Kogan, and Wei Wu, 2021, Common fund flows: Flow hedging and factor pricing, Working paper, University of Pennsylvania.

Ehsani, Sina, and Juhani T Linnainmaa, 2022, Factor momentum and the momentum factor, *The Journal of Finance* 77, 1877–1919.

Fama, Eugene F, and Kenneth R French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Fama, Eugene F, and James D MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy* 81, 607–636.

Gabaix, Xavier, and Ralph SJ Koijen, 2022, In search of the origins of financial fluctuations: The inelastic markets hypothesis, Working paper, Harvard University.

Gârleanu, Nicolae, and Lasse Heje Pedersen, 2022, Active and passive investing: Understanding samuelson's dictum, *Review of Asset Pricing Studies* 12, 389–446.

Hansen, Lars Peter, and Ravi Jagannathan, 1991, Implications of security market data for models of dynamic economies, *Journal of Political Economy* 99, 225–262.

Hasbrouck, Joel, and Duane J Seppi, 2001, Common factors in prices, order flows, and liquidity, *Journal of Financial Economics* 59, 383–411.

Huang, Shiyang, Yang Song, and Hong Xiang, 2021, Noise trading and asset pricing factors, Working paper, The University of Hong Kong.

Kelly, Bryan T, Tobias J Moskowitz, and Seth Pruitt, 2021, Understanding momentum and reversal, *Journal of Financial Economics* 140, 726–743.

Kim, Minsoo, 2020, Fund flows, liquidity, and asset prices, Working paper, University of Melbourne.

Koijen, Ralph SJ, and Motohiro Yogo, 2019, A demand system approach to asset pricing, *Journal of Political Economy* 127, 1475–1515.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2018, Interpreting factor models, *The Journal of Finance* 73, 1183–1223.

Kumar, Praveen, and Duane J Seppi, 1994, Information and index arbitrage, *Journal of Business* 481–509.

Kyle, Albert S, 1985, Continuous auctions and insider trading, *Econometrica* 1315–1335.

Li, Jiacui, and Zihan Lin, 2022, Prices are less elastic at more aggregate levels, Working paper, University of Utah.

Lo, Andrew W, and Jiang Wang, 2000, Trading volume: definitions, data analysis, and implications of portfolio theory, *Review of Financial Studies* 13, 257–300.

Lou, Dong, 2012, A flow-based explanation for return predictability, *Review of Financial Studies* 25, 3457–3489.

Lu, Zhongjin, Steven Malliaris, and Zhongling Qin, 2023, Heterogeneous liquidity providers and night-minus-day return predictability, *Journal of Financial Economics* 148, 175–200.

Merton, Robert C, 1973a, An intertemporal capital asset pricing model, *Econometrica* 867–887.

Merton, Robert C, 1973b, Theory of rational option pricing, *The Bell Journal of Economics and Management Science* 141–183.

Pasquariello, Paolo, and Clara Vega, 2015, Strategic cross-trading in the US stock market, *Review of Finance* 19, 229–282.

Pástor, Ľuboš, and Robert F Stambaugh, 2003, Liquidity risk and expected stock returns, *Journal of Political Economy* 111, 642–685.

Ross, Stephen A, 1976, The arbitrage theory of capital asset pricing, *Journal of Economic Theory* 13, 341–60.

Rostek, Marzena J, and Ji Hee Yoon, 2020, Equilibrium theory of financial markets: Recent developments, Working paper, University of Wisconsin - Madison.

Sims, Christopher A, 2006, Rational inattention: Beyond the linear-quadratic case, *American Economic Review* 96, 158–163.

Train, Kenneth E, 2009, *Discrete choice methods with simulation* (Cambridge university press).

Veldkamp, Laura L, 2011, *Information choice in macroeconomics and finance* (Princeton University Press).

Wang, Jiang, 1993, A model of intertemporal asset prices under asymmetric information, *Review of Economic Studies* 60, 249–282.

Appendix

The appendices provide proofs and additional theoretical results.

A Proofs

In this appendix, I provide proofs omitted in the main text.

A.1 Proof of Proposition 2

First, I write the F-SDF $\Delta M(\mathbf{f})$ in the space spanned by $\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]$ as $\Delta M(\mathbf{f}) = (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])^\top \mathbf{b}$ for some $\mathbf{b} \in \mathbb{R}^N$. By equation (8), I have $\Delta \mathbf{p}(\mathbf{f}) = \text{var}(\mathbf{R}_0)\mathbf{b}$. Therefore, I have $\mathbf{b} = \text{var}(\mathbf{R}_0)^{-1}\Delta \mathbf{p}(\mathbf{f})$ and thus

$$\Delta M(\mathbf{f}) = \Delta \mathbf{p}(\mathbf{f})^\top \text{var}(\mathbf{R}_0)^{-1}(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]). \tag{A.1}$$

Next, I show that linearity Assumption 1 implies the linearity of the F-SDF. Using condition (iii) and equation (A.1), I have, for any $a_1 \in \mathbb{R}$, $a_2 \in \mathbb{R}$, $\mathbf{f}_1 \in \mathbb{R}^N$, and $\mathbf{f}_2 \in \mathbb{R}^N$,

$$\begin{aligned}
\Delta M(a_1\mathbf{f}_1 + a_2\mathbf{f}_2) &= \Delta \mathbf{p}(a_1\mathbf{f}_1 + a_2\mathbf{f}_2)^\top \text{var}(\mathbf{R}_0)^{-1}(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) \\
&= \left(a_1\Delta \mathbf{p}(\mathbf{f}_1) + a_2\Delta \mathbf{p}(\mathbf{f}_2)\right)^\top \text{var}(\mathbf{R}_0)^{-1}(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) \\
&= a_1\Delta M(\mathbf{f}_1) + a_2\Delta M(\mathbf{f}_2).
\end{aligned} \tag{A.2}$$

Then, I define $\mathbf{c}_n$ as an $N \times 1$ vector with only the n-th entry being one and all other entries being zero. Equation (A.2) implies that $\Delta M(\mathbf{f}) = \sum_{n=1}^N f_n \Delta M(\mathbf{c}_n)$. Condition (iii) implies that $\Delta M(\mathbf{c}_n) \in \underline{\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]}$ also resides within the the space spanned by $\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]$, so I can write $\Delta M(\mathbf{c}_n) = \mathbf{y}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])$ for some $\mathbf{y}_n \in \mathbb{R}^N$. Therefore, I have

$$\Delta M(\mathbf{f}) = \sum_{n=1}^N f_n \mathbf{y}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) = (\mathbf{Y}\mathbf{f})^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]), \tag{A.3}$$

where $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$. Equation (19) and Assumption 2 imply that $\text{var}(\mathbf{R}_0)\mathbf{Y}$ is positive definite.

A.2 Proof of Lemma 1

Because the matrix $\text{var}(\mathbf{R}_0)$ has full rank, I carry out Cholesky decomposition and obtain $\text{var}(\mathbf{R}_0) = \mathbf{U}^\top\mathbf{U}$, where $\mathbf{U}$ is an $N \times N$ upper triangular matrix with positive diagonal entries. I then carry out eigenvalue decomposition of the symmetric matrix $\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top$. Because $\text{var}(\mathbf{f})$ has full rank, I obtain

$$\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top\mathbf{G} = \mathbf{G}\mathbf{\Pi}, \tag{A.4}$$

where $\mathbf{\Pi} = \text{diag}(\pi_1, \pi_2, \ldots, \pi_N)$ and $\mathbf{G}$ is an $N \times N$ orthonormal matrix satisfying $\mathbf{G}^\top\mathbf{G} = \mathbf{I}_N$. I then define $\mathbf{B} = \mathbf{U}^{-1}\mathbf{G}$, where the upper triangular matrix $\mathbf{U}$ with positive diagonal entries is by definition invertible.

I now show that the constructed $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N)$ satisfies the conditions (23) and (24). First, I have

$$\mathbf{B}^\top\text{var}(\mathbf{R}_0)\mathbf{B} = \mathbf{G}^\top(\mathbf{U}^\top)^{-1}\mathbf{U}^\top\mathbf{U}\mathbf{U}^{-1}\mathbf{G} = \mathbf{I}_N. \tag{A.5}$$

Second, equation (22) implies $\mathbf{f} = \mathbf{B}\mathbf{q}$, where $\mathbf{q} = (q_1, q_2, \ldots, q_N)^\top$. Therefore,

$$\text{var}(\mathbf{f}) = \mathbf{B}\text{var}(\mathbf{q})\mathbf{B}^\top. \tag{A.6}$$

From (A.4), I have $\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top\mathbf{U}\mathbf{B} = \mathbf{U}\mathbf{B}\mathbf{\Pi}$. Because $\mathbf{U}$ is invertible, I have

$$\text{var}(\mathbf{f})\mathbf{U}^\top\mathbf{U}\mathbf{B} = \mathbf{B}\mathbf{\Pi}. \tag{A.7}$$

Plugging (A.6) into (A.7), I obtain

$$\mathbf{B}\mathbf{\Pi} = \mathbf{B}\text{var}(\mathbf{q})\mathbf{B}^\top\mathbf{U}^\top\mathbf{U}\mathbf{B} = \mathbf{B}\text{var}(\mathbf{q}), \tag{A.8}$$

where I have used $\mathbf{B}^\top \mathbf{U}^\top \mathbf{U}\mathbf{B} = \mathbf{G}^\top \mathbf{G} = \mathbf{I}_N$. Equation (A.8) implies $\text{var}(\mathbf{q}) = \mathbf{\Pi}$.

Finally, I show the uniqueness of the matrix $\mathbf{B}$. Suppose that some other matrix $\tilde{\mathbf{B}}$ satisfies (23) and (24). I define $\tilde{\mathbf{G}} = \mathbf{U}\tilde{\mathbf{B}}$. I then have $\tilde{\mathbf{G}}^\top \tilde{\mathbf{G}} = \mathbf{I}_N$ and $\tilde{\mathbf{G}}^\top \mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top \tilde{\mathbf{G}} = \mathbf{\Pi}$. Assumption 4 requires that $\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top$ has distinct eigenvalues, which implies that matrix $\tilde{\mathbf{G}}$ is unique up to sign for any columns (recall that I have arranged the eigenvalue matrix $\mathbf{\Pi}$ from large to small eigenvalues). Note that $\tilde{\mathbf{B}} = \mathbf{U}^{-1}\tilde{\mathbf{G}}$, where $\mathbf{U}^{-1}$ is also an upper triangular matrix. Therefore, the matrix $\tilde{\mathbf{B}}$ is also unique up to sign for any columns.

A.3 Proof of Theorem 1

I start by showing the following lemma, which allows me to state the equivalent IUF condition under the factors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$.

LEMMA 2. *For any given* $\mathbf{a} \in \mathbb{R}^N$ *and* $\mathbf{d} \in \mathbb{R}^N$, $\mathbf{a}^\top \mathbf{C}\mathbf{d} = 0$ *for all* $\mathbf{C} \in \mathcal{C}$ *if and only if* $\text{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{a}^\top \mathbf{R}_0)\text{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) = 0$ *for all* $n = 1, 2, \ldots, N$.

Proof. Following the proof of Lemma 1 in Appendix A.2, I carry out Cholesky decomposition to obtain $\text{var}(\mathbf{R}_0) = \mathbf{U}^\top \mathbf{U}$, where $\mathbf{U}$ is an $N \times N$ upper triangular matrix. I have $\mathbf{b}_n = \mathbf{U}^{-1}\mathbf{g}_n$ for $n = 1, 2, \ldots, N$, where $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N$ is the full set of eigenvectors of $\text{var}(\mathbf{U}\mathbf{f})$ with distinct eigenvalues $\pi_1 > \pi_2 > \cdots > \pi_N > 0$.

I have that

$$
\begin{aligned}
&\text{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{a}^\top \mathbf{R}_0)\text{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) = 0 \text{ for all } n = 1, 2, \ldots, N \\
\iff\ &\mathbf{a}^\top \text{var}(\mathbf{R}_0)\mathbf{b}_n \mathbf{b}_n^\top \text{var}(\mathbf{R}_0)\mathbf{d} = 0 \text{ for all } n = 1, 2, \ldots, N \\
\iff\ &\mathbf{a}^\top \mathbf{U}^\top \mathbf{g}_n \mathbf{g}_n^\top \mathbf{U}\mathbf{d} = 0 \text{ for all } n = 1, 2, \ldots, N \qquad \text{(A.9)} \\
\iff\ &\mathbf{a}^\top \mathbf{U}^\top \left(\sum_{n=1}^{N} z_n \mathbf{g}_n \mathbf{g}_n^\top \right) \mathbf{U}\mathbf{d} = 0 \text{ for any real numbers } z_1, z_2, \ldots, z_N, \qquad \text{(A.10)}
\end{aligned}
$$

where I have used the Cholesky decomposition of $\text{var}(\mathbf{R}_0)$ in (A.9).

I define the following set of matrices,

$$\mathcal{H} = \left\{ \mathbf{H} \in \mathbb{R}^{N \times N} \,\middle|\, \mathbf{H} = \sum_{n=1}^{N} z_n \mathbf{g}_n \mathbf{g}_n^\top \text{ for some real numbers } z_1, z_2, \ldots, z_N \right\}. \tag{A.11}$$

I claim that

$$\mathcal{H} = \left\{ \mathbf{H} \in \mathbb{R}^{N \times N} \,\middle|\, \mathbf{H}\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top = \mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top\mathbf{H} \right\}. \tag{A.12}$$

It is easy to see that if $\mathbf{H} \in \mathcal{H}$ as defined in (A.11), I have $\mathbf{H} \in \mathcal{H}$ as defined in (A.12) because $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N$ are the eigenvectors of $\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top$. Conversely, if $\mathbf{H} \in \mathcal{H}$ as defined in (A.12), then for any n,

$$\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top(\mathbf{H}\mathbf{g}_n) = (\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top\mathbf{H})\mathbf{g}_n = (\mathbf{H}\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top)\mathbf{g}_n = \pi_n(\mathbf{H}\mathbf{g}_n), \tag{A.13}$$

showing that $\mathbf{H}\mathbf{g}_n$ is also the eigenvector of $\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top$ that corresponds to the eigenvalue π_n. Because the eigenvalues are distinct and the eigenspaces are all one-dimensional, I have $\mathbf{H}\mathbf{g}_n = \mu_n \mathbf{g}_n$ for some μ_n. This shows that $\mathbf{g}_n$ is also an eigenvector of $\mathbf{H}$. Because n is arbitrary for any $1, 2, \ldots, N$, I have $\mathbf{H} = \sum_{n=1}^{N} \mu_n \mathbf{g}_n \mathbf{g}_n^\top$, showing that $\mathbf{H} \in \mathcal{H}$ as defined in (A.11).

Given (A.12), I can equivalently rewrite (A.10) as

$$\mathbf{a}^\top \mathbf{U}^\top \mathbf{H}\mathbf{U}\mathbf{d} = 0 \text{ for any } \mathbf{H} \in \mathcal{H}. \tag{A.14}$$

I define $\mathbf{C} = \mathbf{U}^\top\mathbf{H}\mathbf{U}$. Using the fact that $\mathbf{U}$ is invertible and $\text{var}(\mathbf{R}_0) = \mathbf{U}^\top\mathbf{U}$, I have $\mathbf{C} \in \mathcal{C} \iff \mathbf{H} \in \mathcal{H}$, where $\mathcal{C}$ is defined in (20). Therefore, I can equivalently rewrite (A.14) as $\mathbf{a}^\top\mathbf{C}\mathbf{d} = 0$ for any $\mathbf{C} \in \mathcal{C}$, which completes the proof of Lemma 2. □

Lemma 2 allows me to write the equivalent IUF under factors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$.

Assumption 3*. ***IUF under factors***

For any given portfolio $\mathbf{a} \in \mathbb{R}^N$ *and flow* s*, I construct the portfolio,*

$$\mathbf{d} = (\text{cov}(f_1, s), \text{cov}(f_2, s), \ldots, \text{cov}(f_N, s))^\top . \tag{A.15}$$

If $\text{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{a}^\top \mathbf{R}_0)\text{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) = 0$ *for all* n*, then* $\text{cov}(\mathbf{a}^\top \Delta\mathbf{p}(\mathbf{f}), s) = 0$.

Note that the IUF condition in Assumption 3 is equivalent to Assumption 3* only when Assumption 4 holds. Without Assumption 4, the IUF imposes no restrictions for eigenvectors that correspond to the same eigenvalue. In that case, the factors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$ are also not unique and the IUF condition does not simplify to Assumption 3*. I present the general case in Online Appendix A.

I also prove the following coordinate transformation,

$$\mathbf{d} = \text{cov}(\mathbf{f}, s) = \sum_{n=1}^{N} \mathbf{b}_n \text{cov}(q_n, s), \tag{A.16}$$

where the first equality follows from definition (21) and the second equality uses the factor decomposition (22). This coordinate transformation (A.16) implies that the portfolio $\mathbf{d}$ is a linear combination of the factor portfolios $\mathbf{b}_n$, with weights equal to the covariance between flow s and factor flow q_n. Using (A.16) and orthogonalization condition (23), I have

$$\text{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) = \mathbf{b}_n^\top \text{var}(\mathbf{R}_0) \left(\sum_{m=1}^{N} \mathbf{b}_m \text{cov}(q_m, s) \right) = \text{cov}(q_n, s). \tag{A.17}$$

That is, the covariance between the return on the factor portfolio $\mathbf{b}_n$ and the return on the portfolio $\mathbf{d}$ equals the covariance between the corresponding factor flow q_n and flow s.

Given these observations, I can proceed to prove the theorem. First, I show that the F-SDF form (25) satisfies the IUF in Assumption 3*. (It is evident that the form (25) is a special case of the form (18), so automatically satisfies all restriction in Proposition 2.) I

derive the price impact under the F-SDF form (25) as

$$\Delta \mathbf{p}(\mathbf{f}) = \mathbb{E}[\mathbf{R}_0 \Delta M(\mathbf{f})] = \mathrm{var}(\mathbf{R}_0) \sum_{n=1}^{N} \lambda_n q_n \mathbf{b}_n. \tag{A.18}$$

I project any portfolio $\mathbf{a} \in \mathbb{R}^N$ onto the factor portfolios $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$ and obtain

$$\mathbf{a} = \sum_{n=1}^{N} x_n \mathbf{b}_n \tag{A.19}$$

for some real numbers $x_1, x_2, \ldots, x_N$. The condition of the IUF in Assumption 3* is

$$\begin{aligned} 0 &= \mathrm{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{a}^\top \mathbf{R}_0) \mathrm{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) \\ &= \left(\sum_{m=1}^{N} x_m \mathbf{b}_m \right)^\top \mathrm{var}(\mathbf{R}_0) \mathbf{b}_n \mathrm{cov}(q_n, s) = x_n \mathrm{cov}(q_n, s), \end{aligned} \tag{A.20}$$

for any $n = 1, 2, \ldots, N$, where I use equations (23), (A.17), and (A.19).

I have

$$\mathbf{a}^\top \Delta \mathbf{p}(\mathbf{f}) = \left(\sum_{m=1}^{N} x_m \mathbf{b}_m \right)^\top \mathrm{var}(\mathbf{R}_0) \left(\sum_{n=1}^{N} \lambda_n q_n \mathbf{b}_n \right) = \sum_{n=1}^{N} x_n \lambda_n q_n, \tag{A.21}$$

where I use equations (23), (A.18), and (A.19). Therefore, I have

$$\mathrm{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\mathbf{f}), s) = \sum_{n=1}^{N} x_n \mathrm{cov}(q_n, s) \lambda_n. \tag{A.22}$$

Note that the IUF's assumption implies that $x_n \mathrm{cov}(q_n, s) = 0$ for any $n = 1, 2, \ldots, N$ in (A.20). Therefore, by equation (A.22), I have $\mathrm{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\mathbf{f}), s) = 0$, thereby showing that the IUF holds. Moreover, equation (A.18) implies that

$$\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f}) = \mathbf{q}^\top \mathbf{B}^\top \mathrm{var}(\mathbf{R}_0) \mathbf{B} \mathbf{\Lambda} \mathbf{q} = \mathbf{q}^\top \mathbf{\Lambda} \mathbf{q}, \tag{A.23}$$

where $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N)$, $\mathbf{q} = (q_1, q_2, \ldots, q_N)^\top$, and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$. Because $\mathbf{\Lambda}$ is positive definite, Assumption 2 holds.

Second, I show that, if the F-SDF in (18) satisfies the IUF, the F-SDF can be written in the form (25). Using equation (22), I simplify (18) as

$$\begin{aligned}\Delta M(\mathbf{f}) &= \sum_{n=1}^{N}\sum_{m=1}^{N} q_m b_{n,m} \mathbf{g}_n^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) \\ &= \sum_{m=1}^{N} q_m \sum_{n=1}^{N} b_{n,m} \mathbf{g}_n^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) = \sum_{m=1}^{N} q_m \mathbf{h}_m^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]),\end{aligned} \tag{A.24}$$

where I define $\mathbf{h}_m = \sum_{n=1}^{N} b_{n,m}\mathbf{g}_n$. Therefore, the price impact under the F-SDF (A.24) is

$$\Delta \mathbf{p}(\mathbf{f}) = \mathbb{E}[\mathbf{R}_0 \Delta M(\mathbf{f})] = \text{var}(\mathbf{R}_0) \sum_{m=1}^{N} q_m \mathbf{h}_m. \tag{A.25}$$

I use the IUF for portfolio $\mathbf{a} = \mathbf{b}_l$ and flow $s = q_m$ for any $l \neq m$. I have from (A.16) and orthogonalization condition (24),

$$\mathbf{d} = \sum_{n=1}^{N} \mathbf{b}_n \text{cov}(q_n, q_m) = \pi_m \mathbf{b}_m. \tag{A.26}$$

Thus, I have $\text{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{a}^\top \mathbf{R}_0)\text{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) = 0$ for any $n = 1, 2, \ldots, N$, because of orthogonalization condition (23). Thus, the IUF implies that

$$0 = \text{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\mathbf{f}), s) = \text{cov}\left(\mathbf{b}_l^\top \text{var}(\mathbf{R}_0) \sum_{n=1}^{N} q_n \mathbf{h}_n, q_m\right) = \pi_m \mathbf{b}_l^\top \text{var}(\mathbf{R}_0) \mathbf{h}_m. \tag{A.27}$$

Thus, for any given $m = 1, 2, \ldots, N$, I can choose arbitrary $l \neq m$, such that

$$\mathbf{b}_l^\top \text{var}(\mathbf{R}_0) \mathbf{h}_m = 0. \tag{A.28}$$

I project $\mathbf{h}_m$ onto the factor portfolios $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$ and obtain $\mathbf{h}_m = \sum_{n=1}^{N} \theta_{m,n} \mathbf{b}_n$. Using the IUF condition (A.28) and orthogonalization condition (23), I have $\theta_{m,n} = 0$ for any

$m \neq n$. Therefore, I can rewrite the F-SDF form (A.24) as

$$\Delta M(\mathbf{f}) = \sum_{n=1}^{N} \theta_{n,n} q_n \mathbf{b}_n^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]). \tag{A.29}$$

I rename $\theta_{n,n} = \lambda_n$ and obtain the form (25). Lastly, using (A.23) and Assumption 2, I have $\lambda_n > 0$ for all n. The proof is complete.

A.4 Proof of Theorem 2

I first show the sufficiency direction. From Theorem 1 and Proposition 5, if the orthogonalized model (32) satisfies the IUF, there exists some $N \times N$ invertible matrix $\mathbf{O}$, such that $\mathbf{O}^{-1}\text{diag}(\tilde{\lambda}_{1,1}, \tilde{\lambda}_{2,2}, \ldots, \tilde{\lambda}_{N,N})\mathbf{O}$ is a diagonal matrix for any free parameters $\tilde{\lambda}_{1,1}, \tilde{\lambda}_{2,2}, \ldots, \tilde{\lambda}_{N,N}$. Denote the (n,m)-th entry of $\mathbf{O}^{-1}$ as $x_{n,m}$ and the (n,m)-th entry of $\mathbf{O}$ as $y_{n,m}$. I then know that $x_{n,m}y_{m,l} = 0$ for any m and $n \neq l$. Suppose that there exist some m and $l \neq l'$, such that $y_{m,l} \neq 0$ and $y_{m,l'} \neq 0$. I then know that $x_{n,m} = 0$ for any $n = 1, 2, \ldots, N$. This contradicts the fact that $\mathbf{O}$ is invertible. Therefore, any row of $\mathbf{O}$ has exactly one non-zero element. Together with the fact that $\mathbf{O}$ is invertible, any column of $\mathbf{O}$ also has exactly one non-zero element. In other words, $\mathbf{O}$ is just a scaling and reordering matrix. By Proposition 5, I have $\text{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ and $\text{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ for any $n \neq m$.

To show (36), note that I have, by (31) and $\text{cov}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0) = 0$ for $n \neq l$,

$$\tilde{\mathbf{b}}_l^\top \Delta\mathbf{p}(\mathbf{f}) = \sum_{n=1}^{N}\sum_{m=1}^{N} \tilde{\lambda}_{n,m}\tilde{q}_m \text{cov}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0) = \sum_{m=1}^{N} \tilde{\lambda}_{l,m}\tilde{q}_m \text{var}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0). \tag{A.30}$$

I then have

$$\frac{\partial \tilde{\mathbf{b}}_l^\top \Delta\mathbf{p}(\mathbf{f})}{\partial \tilde{q}_l} = \tilde{\lambda}_{l,l}\text{var}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0). \tag{A.31}$$

Similarly, I have, by (32),

$$\tilde{\mathbf{b}}_l^\top \Delta\bar{\mathbf{p}}(\mathbf{f}) = \sum_{n=1}^{N} \tilde{\lambda}_{n,n}\tilde{q}_n \text{cov}(\tilde{\mathbf{b}}_l \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0) = \tilde{\lambda}_{l,l}\tilde{q}_l \text{var}(\tilde{\mathbf{b}}_l \mathbf{R}_0), \tag{A.32}$$

and

$$\frac{\partial \tilde{\mathbf{b}}_l^\top \Delta \bar{\mathbf{p}}(\mathbf{f})}{\partial \tilde{q}_l} = \tilde{\lambda}_{l,l} \text{var}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0). \tag{A.33}$$

This shows that (36) holds.

Under the price impact model (31), the expected compensation for risk is

$$\mathbb{E}[\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f})] = \sum_{n=1}^{N} \sum_{m=1}^{N} \tilde{\lambda}_{n,m} \text{var}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0) \text{cov}(\tilde{q}_m, \tilde{q}_n) = \sum_{n=1}^{N} \tilde{\lambda}_{n,n} \text{var}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0) \text{var}(\tilde{q}_n), \tag{A.34}$$

where the first equality uses $\text{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$, and the second equality uses $\text{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ for any $n \neq m$. Therefore, the expected compensation $\mathbb{E}[\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f})]$ under models (31) and (32) are the same for any N^2 parameters $\tilde{\lambda}_{n,m}$. By equation (35), the expected utility $\mathbb{E}[u(W(\mathbf{f}))]$ under models (31) and (32) are also the same.

Next, I show the necessity direction. I have, by (31),

$$\tilde{\mathbf{b}}_l^\top \Delta \mathbf{p}(\mathbf{f}) = \sum_{n=1}^{N} \sum_{m=1}^{N} \tilde{\lambda}_{n,m} \tilde{q}_m \text{cov}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0). \tag{A.35}$$

I then have

$$\frac{\partial \tilde{\mathbf{b}}_l^\top \Delta \mathbf{p}(\mathbf{f})}{\partial \tilde{q}_l} = \sum_{n=1}^{N} \tilde{\lambda}_{n,l} \text{cov}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0). \tag{A.36}$$

Similarly, I have, by (32),

$$\tilde{\mathbf{b}}_l^\top \Delta \bar{\mathbf{p}}(\mathbf{f}) = \sum_{n=1}^{N} \tilde{\lambda}_{n,n} \tilde{q}_n \text{cov}(\tilde{\mathbf{b}}_l \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0), \tag{A.37}$$

and

$$\frac{\partial \tilde{\mathbf{b}}_l^\top \Delta \bar{\mathbf{p}}(\mathbf{f})}{\partial \tilde{q}_l} = \tilde{\lambda}_{l,l} \text{var}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0). \tag{A.38}$$

Because (36) holds for any parameters $\tilde{\lambda}_{n,m}$, I have $\text{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ for any $n \neq m$.

Under the price impact model (31), the expected compensation is

$$\mathbb{E}[\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f})] = \sum_{n=1}^{N}\sum_{m=1}^{N} \tilde{\lambda}_{n,m} \mathrm{var}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0) \mathrm{cov}\left(\tilde{q}_n, \tilde{q}_m\right), \tag{A.39}$$

where I use $\mathrm{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ for any $n \neq m$. By (35), if the expected utility $\mathbb{E}[u(W(\mathbf{f}))]$ remains invariant under models (31) and (32), the expected compensation $\mathbb{E}[\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f})]$ also remains invariant. Therefore, $\mathrm{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ for any $n \neq m$. By Theorem 1, $\mathrm{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ and $\mathrm{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ for any $n \neq m$ imply that the orthogonalized model (32) satisfies the IUF for any parameters $\tilde{\lambda}_{1,1}, \tilde{\lambda}_{2,2}, \ldots, \tilde{\lambda}_{N,N}$.

A.5 Proof of Proposition 3

I write the F-SDF $\Delta M(\mathbf{f})$ in the space spanned by $\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]$ as $\Delta M^*(\mathbf{f}) = (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])^\top \mathbf{b}$ for some $\mathbf{b} \in \mathbb{R}^N$. By equation (8), I have $\Delta \mathbf{p}(\mathbf{f}) = \mathrm{var}(\mathbf{R}_0)\mathbf{b}$. Therefore, I have $\mathbf{b} = \mathrm{var}(\mathbf{R}_0)^{-1}\Delta \mathbf{p}(\mathbf{f})$ and thus

$$\Delta M^*(\mathbf{f}) = (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])^\top \mathrm{var}(\mathbf{R}_0)^{-1} \Delta \mathbf{p}(\mathbf{f}). \tag{A.40}$$

The variance of $\Delta M^*(\mathbf{f})$ is $\mathrm{var}(\Delta M^*(\mathbf{f})) = \Delta \mathbf{p}(\mathbf{f})^\top \mathrm{var}(\mathbf{R}_0)^{-1} \Delta \mathbf{p}(\mathbf{f})$.

I claim that

$$\mathrm{var}(\Delta M^*(\mathbf{f})) \leq \max_{\mathbf{c} \in \mathbb{R}^N} \frac{\Delta \mathbf{p}(\mathbf{f})^\top \mathbf{c}\mathbf{c}^\top \Delta \mathbf{p}(\mathbf{f})}{\mathrm{var}(\mathbf{c}^\top \mathbf{R}_0)}. \tag{A.41}$$

I choose portfolio $\mathbf{c} = \mathrm{var}(\mathbf{R}_0)^{-1}\Delta \mathbf{p}(\mathbf{f})$. Therefore, I have

$$\frac{\Delta \mathbf{p}(\mathbf{f})^\top \mathbf{c}\mathbf{c}^\top \Delta \mathbf{p}(\mathbf{f})}{\mathrm{var}(\mathbf{c}^\top \mathbf{R}_0)} = \frac{(\Delta \mathbf{p}(\mathbf{f})^\top \mathrm{var}(\mathbf{R}_0)^{-1} \Delta \mathbf{p}(\mathbf{f}))^2}{\mathbf{c}^\top \mathrm{var}(\mathbf{R}_0)\mathbf{c}} = \Delta \mathbf{p}(\mathbf{f})^\top \mathrm{var}(\mathbf{R}_0)^{-1} \Delta \mathbf{p}(\mathbf{f}). \tag{A.42}$$

Therefore, I have proved (A.41). Combining with inequality (38) in the main text, I have proved the flow-based Hansen-Jagannathan bound.

A.6 Proof of Proposition 4

By Proposition 3, condition (43) implies

$$\max_{\|\mathbf{f}\|=1} \Delta\mathbf{p}(\mathbf{f})^\top \text{var}(\mathbf{R}_0)^{-1} \Delta\mathbf{p}(\mathbf{f}) \leq H^2. \tag{A.43}$$

Under the N-factor model (26), the squared maximum price impact ratio is

$$\Delta\mathbf{p}(\mathbf{f})^\top \text{var}(\mathbf{R}_0)^{-1} \Delta\mathbf{p}(\mathbf{f}) = \sum_{n=1}^{N} \lambda_n q_n \mathbf{b}_n^\top \text{var}(\mathbf{R}_0) \sum_{n=1}^{N} \lambda_n q_n \mathbf{b}_n = \sum_{n=1}^{N} \lambda_n^2 q_n^2, \tag{A.44}$$

where I use condition (23). I consider the flow $\mathbf{f} = \sum_{n=1}^{N} q_n \mathbf{b}_n$, with $q_m = 1/\|\mathbf{b}_m\|$ for some specific m and $q_m = 0$ for $m \neq n$. Clearly, $\|\mathbf{f}\| = 1$, so by (A.43) and (A.44), I have $|\lambda_m| \leq H\|\mathbf{b}_m\|$. That is, all λ_n are uniformly bounded for any F-SDF that satisfies (43).

I can simplify the pricing error of the price impact as

$$\Delta\mathbf{p}(\mathbf{f}) - \Delta\check{\mathbf{p}}(\mathbf{f}) = \text{var}(\mathbf{R}_0) \sum_{n=K+1}^{N} \lambda_n q_n \mathbf{b}_n = \sum_{n=K+1}^{N} \mathbf{c}_n q_n, \tag{A.45}$$

where I define the $N \times 1$ vector $\mathbf{c}_n = \text{var}(\mathbf{R}_0)\lambda_n \mathbf{b}_n$. Because λ_n are bounded, each element of every $\mathbf{c}_n$ is also uniformly bounded for any F-SDF that satisfies (43). I have

$$\|\Delta\mathbf{p}(\mathbf{f}) - \Delta\check{\mathbf{p}}(\mathbf{f})\|^2 = \sum_{n=1}^{N} \left(\sum_{k=K+1}^{N} c_{n,k} q_k \right)^2. \tag{A.46}$$

Using the fact that q_k are uncorrelated with each other, I have

$$\begin{aligned}\mathbb{E}[\|\Delta\mathbf{p}(\mathbf{f}) - \Delta\check{\mathbf{p}}(\mathbf{f})\|^2] &= \sum_{n=1}^{N}\mathbb{E}\left(\sum_{k=K+1}^{N} c_{n,k}q_k\right)^2 \\ &= \sum_{n=1}^{N}\sum_{k=K+1}^{N} c_{n,k}^2\mathbb{E}[q_k^2] + 2\sum_{n=1}^{N}\sum_{K+1\le i<j\le N} c_{n,i}c_{n,j}\mathbb{E}[q_i]\mathbb{E}[q_j] \\ &\le \sum_{n=1}^{N}\sum_{k=K+1}^{N} c_{n,k}^2\mathbb{E}[q_k^2] + \sum_{n=1}^{N}\sum_{K+1\le i<j\le N}\left(c_{n,i}^2\mathbb{E}[q_i]^2 + c_{n,j}^2\mathbb{E}[q_j]^2\right).\end{aligned} \tag{A.47}$$

Because all $c_{n,k}$ are uniformly bounded for any F-SDF that satisfies (43), if $\sum_{n=K+1}^{N}\mathbb{E}[q_k^2]$ tends to zero, then $\mathbb{E}[\|\Delta\mathbf{p}(\mathbf{f}) - \Delta\check{\mathbf{p}}(\mathbf{f})\|^2]$ tends to zero.

B F-SDF and Price Impacts Under Alternative Factors

In practice, one may not want to use the factor portfolios $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N)$ constructed in Lemma 1 to characterize the F-SDF and price impacts. Instead, one might want to choose a set of portfolios $\tilde{\mathbf{B}} = (\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N)$ that have more explicit economic interpretations. For example, $\tilde{\mathbf{b}}_1$ could be the Fama-French high-minus-low (HML) portfolio, and $\tilde{\mathbf{b}}_2$ could be the small-minus-big (SMB) portfolio. In this case, the factor flow $\tilde{\mathbf{q}} = (\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_N)^\top$ defined using $\mathbf{f} = \tilde{\mathbf{B}}\tilde{\mathbf{q}}$ also has more explicit interpretations. Using the same example, $\tilde{q}_1$ is the HML factor flow, and $\tilde{q}_2$ is the SMB factor flow.

Because these factors generally do not have uncorrelated flows and uncorrelated fundamental returns, they still have cross-impacts between each other. The IUF implies a structural restriction on the cross-impacts between the factors, which is represented by an $N \times N$ price-of-flow-induced-risk matrix $\tilde{\mathbf{\Lambda}} = \{\tilde{\lambda}_{n,m}\}$. When these factors are properly chosen to have both uncorrelated flows and uncorrelated fundamental returns, $\tilde{\mathbf{\Lambda}} = \{\tilde{\lambda}_{n,m}\}$ simplifies to a diagonal matrix as in Theorem 1.

PROPOSITION 5. *Under alternative factors, the F-SDF in Theorem 1 becomes*

$$\Delta M(\mathbf{f}) = \sum_{n=1}^{N}\sum_{m=1}^{N} \tilde{\lambda}_{n,m}\tilde{q}_m \tilde{\mathbf{b}}_n^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) = (\tilde{\mathbf{\Lambda}}\tilde{\mathbf{q}})^\top \tilde{\mathbf{B}}^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]), \tag{B.1}$$

and the price impact model becomes

$$\Delta \mathbf{p}(\mathbf{f}) = \sum_{n=1}^{N}\sum_{m=1}^{N} \tilde{\lambda}_{n,m}\tilde{q}_m \mathrm{cov}(\mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0). \tag{B.2}$$

The price-of-flow-induced-risk matrix $\tilde{\mathbf{\Lambda}}$ *satisfies, for some* $N \times N$ *invertible matrix* $\mathbf{O}$*,*

$$\mathbf{O}^{-1}\tilde{\mathbf{\Lambda}}\mathbf{O} = \mathbf{\Lambda}, \text{ where } \mathbf{\Lambda} \text{ is diagonal and positive definite,} \tag{B.3}$$

$$\mathbf{O}^\top \tilde{\mathbf{B}}^\top \mathrm{var}(\mathbf{R}_0)\tilde{\mathbf{B}}\mathbf{O} = \mathbf{I}_N, \tag{B.4}$$

$$\mathbf{O}\mathbf{\Pi}\mathbf{O}^\top = \mathrm{var}(\tilde{\mathbf{q}}), \text{ where } \mathbf{\Pi} \text{ is diagonal and positive definite.} \tag{B.5}$$

Proof. Note that there exists some $N \times N$ invertible matrix $\mathbf{O}$ such that $\tilde{\mathbf{B}}\mathbf{O} = \mathbf{B}$ and $\mathbf{O}\mathbf{q} = \tilde{\mathbf{q}}$, where $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N)$ and $\mathbf{q} = (q_1, q_2, \ldots, q_N)^\top$ are the factor portfolios and flows in equation (25). By (B.1), I have

$$\Delta M(\mathbf{f}) = (\mathbf{O}^{-1}\tilde{\mathbf{\Lambda}}\mathbf{O}\mathbf{q})^\top \mathbf{B}^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]). \tag{B.6}$$

By (25), the above equation implies that $\mathbf{O}^{-1}\tilde{\mathbf{\Lambda}}\mathbf{O} = \mathbf{\Lambda}$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ is the diagonal price-of-flow-induced-risk matrix. Because $\mathbf{B}$ and $\mathbf{q}$ satisfy Lemma 1, conditions (23) and (24) translate into (B.4) and (B.5). □

C Static Quadratic-Normal Price Impact Model

In this appendix, I show that the static quadratic-normal price impact model is a special case of my factor model (26) with an additional restriction $\lambda_1 = \lambda_2 = \cdots = \lambda_N$.

Consider a two-period economy $t = 0$ and $t = 1$. There is a mass μ of infinitesimal

arbitrageurs. There are N assets in the economy, indexed by $n = 1, 2, \ldots, N$. The total fixed supply of assets is an $N \times 1$ vector $\mathbf{S}$, where the unit of supply is the number of shares. These assets are held equally by all arbitrageurs. A risk-free bond is in perfectly elastic supply at a gross interest rate $R_F > 1$. The N assets have payoff $\mathbf{X}$ at time $t = 1$, which is an $N \times 1$ vector of random variables. At time 0, the flow into asset n is h_n, grouped as vector $\mathbf{h} = (h_1, h_2, \ldots, h_N)$. Unlike in the baseline setup, the unit of flow is expressed in the number of shares, not dollar value. I later translate the flow in dollar value. With flow $\mathbf{h}$, each arbitrageur now holds $(\mathbf{S} - \mathbf{h})/\mu$ shares of assets, and I denote the time-0 price of assets as the $N \times 1$ vector $\mathbf{P}(\mathbf{h})$.

The setup so far is equivalent to my baseline setup in Section 3. I now introduce the quadratic-normal assumptions:

1. Each arbitrageur has a CARA utility with parameter γ.

2. Payoff $\mathbf{X}$ is normally distributed with mean $\mathbf{u}$ and variance $\text{var}(\mathbf{X})$.

In equilibrium, the arbitrageurs' optimality condition implies

$$-\mathbf{h}/\mu = \arg\max_{\mathbf{y}} \mathbb{E}[-\exp(-\gamma W(\mathbf{y}))], \tag{C.1}$$

where the time-1 wealth of each arbitrageur is

$$W(\mathbf{y}) = \mathbf{S}^\top \mathbf{X}/\mu + \mathbf{y}^\top (\mathbf{X} - \mathbf{P}(\mathbf{h}) R_F). \tag{C.2}$$

Standard calculation implies that

$$\mathbf{P}(\mathbf{h}) = \frac{\mathbf{u}}{R_F} - \frac{\gamma}{\mu R_F} \text{var}(\mathbf{X})(\mathbf{S} - \mathbf{h}). \tag{C.3}$$

Therefore, the price change is

$$\mathbf{P}(\mathbf{h}) - \mathbf{P}(\mathbf{0}) = \frac{\gamma}{\mu R_F}\text{var}(\mathbf{X})\mathbf{h}. \tag{C.4}$$

I define price impact as

$$\Delta\mathbf{p}(\mathbf{h}) = \left(\frac{P_1(\mathbf{h}) - P_1(\mathbf{0})}{P_1(\mathbf{0})}, \frac{P_2(\mathbf{h}) - P_2(\mathbf{0})}{P_2(\mathbf{0})}, \ldots, \frac{P_N(\mathbf{h}) - P_N(\mathbf{0})}{P_N(\mathbf{0})}\right)^\top, \tag{C.5}$$

and fundamental return as

$$\mathbf{R}_0 = \left(\frac{X_1}{P_1(\mathbf{0})}, \frac{X_2}{P_2(\mathbf{0})}, \ldots, \frac{X_N}{P_N(\mathbf{0})}\right)^\top. \tag{C.6}$$

Using the fact that

$$\text{var}(\mathbf{X}) = \text{diag}(P_1(\mathbf{0}), P_2(\mathbf{0}), \ldots, P_N(\mathbf{0}))\text{var}(\mathbf{R}_0)\text{diag}(P_1(\mathbf{0}), P_2(\mathbf{0}), \ldots, P_N(\mathbf{0})), \tag{C.7}$$

I have

$$\Delta\mathbf{p}(\mathbf{h}) = \frac{\gamma}{\mu R_F}\text{var}(\mathbf{R}_0)\text{diag}(P_1(\mathbf{0}), P_2(\mathbf{0}), \ldots, P_N(\mathbf{0}))\mathbf{h}. \tag{C.8}$$

At this stage, I define the flow in dollar value as

$$\mathbf{f} = (P_1(\mathbf{0})h_1, P_2(\mathbf{0})h_2, \ldots, P_N(\mathbf{0})h_N)^\top = \text{diag}(P_1(\mathbf{0}), P_2(\mathbf{0}), \ldots, P_N(\mathbf{0}))\mathbf{h}. \tag{C.9}$$

Using equations (C.8) and (C.9), I obtain

$$\Delta\mathbf{p}(\mathbf{f}) = \frac{\gamma}{\mu R_F}\text{var}(\mathbf{R}_0)\mathbf{f}. \tag{C.10}$$

Using the factor portfolios and flows constructed in Lemma 1, I have $\mathbf{f} = \sum_{n=1}^{N} q_n \mathbf{b}_n$. There-

fore, the quadratic-normal price impact model is

$$\Delta \mathbf{p}(\mathbf{f}) = \text{var}(\mathbf{R}_0) \sum_{n=1}^{N} \frac{\gamma}{\mu R_F} q_n \mathbf{b}_n. \tag{C.11}$$

Comparing (C.11) with my N-factor price impact model,

$$\Delta \mathbf{p}(\mathbf{f}) = \text{var}(\mathbf{R}_0) \sum_{n=1}^{N} \lambda_n q_n \mathbf{b}_n, \tag{C.12}$$

one sees that the quadratic-normal framework imposes stronger restrictions on the price of flow-induced risk than my model and requires, additionally, $\lambda_1 = \lambda_2 = \cdots = \lambda_N$.

Online Appendix of

"Flow-Based Arbitrage Pricing Theory"

The online appendix provides additional theoretical results omitted in the paper.

A General Theory with Duplicate Eigenvalues

In this appendix, I provide the general theory of the F-SDF and price impacts without imposing regularity Assumption 4 in the main text. That is, I allow the matrix var($\mathbf{Uf}$) to have possibly duplicate eigenvalues.

A.1 Factor Model of F-SDF and Price Impacts

To construct the form of F-SDF under the general case, I conduct eigenvalue decomposition to obtain

$$\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top\mathbf{G} = \mathbf{G}\mathbf{\Pi}, \tag{OA.1}$$

where the eigenvalue matrix is

$$\mathbf{\Pi} = \begin{pmatrix} \pi_1\mathbf{I}_{r_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \pi_2\mathbf{I}_{r_2} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \pi_J\mathbf{I}_{r_J} \end{pmatrix} \tag{OA.2}$$

with $\pi_1 > \pi_2 > \cdots > \pi_J > 0$. The matrix $\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top$ has J distinct positive eigenvalues, with each eigenvalue π_j having r_j degrees of duplication for $j = 1, 2, \ldots, J$. With duplicate eigenvalues, eigenvectors $\mathbf{G}$ are generally not unique. I arbitrarily pick one and construct the corresponding factors $\mathbf{B} = \mathbf{U}^{-1}\mathbf{G}$ following the proof of Lemma 1 in the main text.

Using the N factors $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N)$ and their corresponding flows $\mathbf{q} = (q_1, q_2, \ldots, q_N)^\top$, I now generalize Theorem 1 of the main text.

Theorem O.1. *The F-SDF satisfies all restrictions in Proposition 2 and the IUF Assumption 3* ***if and only if*** *it can be written as*

$$\Delta M(\mathbf{f}) = (\mathbf{\Lambda q})^\top \mathbf{B}^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]), \tag{OA.3}$$

where the price-of-flow-induced-risk matrix $\mathbf{\Lambda}$ *is an* $N \times N$ *block-diagonal matrix*

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Psi}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Psi}_2 & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Psi}_J \end{pmatrix}, \tag{OA.4}$$

where each $\mathbf{\Psi}_j$ *is an* $r_j \times r_j$ *positive definite matrix for* $j = 1, 2, \ldots, J$. *The corresponding price impact model is*

$$\Delta \mathbf{p}(\mathbf{f}) = \mathbb{E}[\mathbf{R}_0 \Delta M(\mathbf{f})] = \text{var}(\mathbf{R}_0)\mathbf{B\Lambda q}. \tag{OA.5}$$

Proof. First, I show that the form (OA.3) satisfies the IUF. That is, for any portfolio $\mathbf{a} \in \mathbb{R}^N$, flow s, and the corresponding portfolio $\mathbf{d}$ that satisfy $\mathbf{a}^\top \mathbf{C d} = 0$ for all $\mathbf{C} \in \mathcal{C}$, the goal is to show $\text{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\mathbf{f}), s) = 0$.

I project portfolio $\mathbf{a} \in \mathbb{R}^N$ onto the factors $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$,

$$\mathbf{a} = \sum_{n=1}^{N} x_n \mathbf{b}_n = \mathbf{Bx}, \tag{OA.6}$$

with the $N \times 1$ vector $\mathbf{x} = (x_1, x_2, \ldots, x_N)^\top$. I can therefore simplify

$$\text{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\mathbf{f}), s) = \mathbf{a}^\top \text{var}(\mathbf{R}_0)\mathbf{B\Lambda} \text{cov}(\mathbf{q}, s) = \mathbf{x}^\top \mathbf{\Lambda} \text{cov}(\mathbf{q}, s), \tag{OA.7}$$

because of the orthogonalization condition $\mathbf{B}^\top \text{var}(\mathbf{R}_0)\mathbf{B} = \mathbf{I}_N$. Note that $\mathbf{\Lambda}$ is a block-diagonal matrix of the form (OA.4). Therefore, to show $\text{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\mathbf{f}), s) = 0$, it is sufficient

to show that $x_l \text{cov}(q_m, s) = 0$ for all l and m such that the (l,l)-th and (m,m)-th elements of matrix $\mathbf{\Pi}$ in (OA.2) correspond to the same eigenvalue.

Using the coordination transformation (A.16) in the main text and equation (OA.6), I simplify the IUF condition as

$$0 = \mathbf{a}^\top \mathbf{C}\mathbf{d} = \mathbf{x}^\top \mathbf{B}^\top \mathbf{C}\mathbf{B}\text{cov}(\mathbf{q}, s). \tag{OA.8}$$

I define the set

$$\mathcal{H} = \left\{ \mathbf{H} \in \mathbb{R}^{N \times N} \,\middle|\, \mathbf{H}\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top = \mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top \mathbf{H} \right\}. \tag{OA.9}$$

Recall that I define the matrix $\mathbf{B} = \mathbf{U}^{-1}\mathbf{G}$ from a given of set of eigenvectors $\mathbf{G}$ of matrix $\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top$ (see equation (OA.1)). Using the transformation $\mathbf{C} = \mathbf{U}^\top \mathbf{H}\mathbf{U}$, I have that $\mathbf{a}^\top \mathbf{C}\mathbf{d} = 0$ for all $\mathbf{C} \in \mathcal{C}$ if and only if $\mathbf{x}^\top \mathbf{G}^\top \mathbf{H}\mathbf{G}\text{cov}(\mathbf{q}, s) = 0$ for all $\mathbf{H} \in \mathcal{H}$.

I define $\mathbf{O}$ as an $N \times N$ matrix with only the (l,m)-th element being one and all other elements being zero. I define the matrix $\tilde{\mathbf{H}} = \mathbf{G}\mathbf{O}\mathbf{G}^\top$. Using (OA.1), I have

$$\tilde{\mathbf{H}}\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top = \mathbf{G}\mathbf{O}\mathbf{G}^\top\mathbf{G}\mathbf{\Pi}\mathbf{G}^\top = \mathbf{G}\mathbf{O}\mathbf{\Pi}\mathbf{G}^\top = \mathbf{G}\mathbf{\Pi}\mathbf{O}\mathbf{G}^\top = \mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top\tilde{\mathbf{H}}, \tag{OA.10}$$

where the key step $\mathbf{\Pi}\mathbf{O} = \mathbf{O}\mathbf{\Pi}$ uses the fact that the (l,l)-th and (m,m)-th elements of matrix $\mathbf{\Pi}$ in (OA.2) correspond to the same eigenvalue. Therefore, $\tilde{\mathbf{H}} \in \mathcal{H}$ and

$$0 = \mathbf{x}^\top \mathbf{G}^\top \tilde{\mathbf{H}}\mathbf{G}\text{cov}(\mathbf{q}, s) = \mathbf{x}^\top \mathbf{O}\text{cov}(\mathbf{q}, s) = x_l \text{cov}(q_m, s). \tag{OA.11}$$

Because l and m are arbitrary as long as they correspond to the same eigenvalue in $\mathbf{\Pi}$, I have that $\text{cov}(\mathbf{a}^\top \Delta\mathbf{p}(\mathbf{f}), s) = 0$, implying that the IUF holds.

Moreover, equation (OA.5) implies that

$$\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f}) = \mathbf{q}^\top \mathbf{B}^\top \text{var}(\mathbf{R}_0)\mathbf{B}\mathbf{\Lambda}\mathbf{q} = \mathbf{q}^\top \mathbf{\Lambda}\mathbf{q}. \tag{OA.12}$$

Because $\mathbf{\Lambda}$ is positive definite, $\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f}) > 0$ for any $\mathbf{f} \neq \mathbf{0}$.

Second, I show that the IUF implies that the F-SDF can be written in the canonical form (OA.3). Following the proof of Theorem 1 in the main text, I write the F-SDF as

$$\Delta M(\mathbf{f}) = \sum_{n=1}^{N} q_n \mathbf{h}_n^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) \tag{OA.13}$$

for some $\mathbf{h}_n \in \mathbb{R}^N$ and derive the corresponding price impact model as

$$\Delta\mathbf{p}(\mathbf{f}) = \text{var}(\mathbf{R}_0) \sum_{n=1}^{N} q_n \mathbf{h}_n. \tag{OA.14}$$

I use the IUF for portfolio $\mathbf{a} = \mathbf{b}_l$ and flow $s = q_m$, where the (l,l)-th and (m,m)-th elements of matrix $\mathbf{\Pi}$ in (OA.2) correspond to distinct eigenvalues. Using the coordinate transformation (A.16) in the main text, I have

$$\mathbf{d} = \sum_{n=1}^{N} \mathbf{b}_n \text{cov}(q_n, q_m) = \text{var}(q_m)\mathbf{b}_m. \tag{OA.15}$$

I next show that $\mathbf{b}_l^\top \mathbf{C} \mathbf{b}_m = 0$ for all $\mathbf{C} \in \mathcal{C}$, where $\mathcal{C}$ is defined in the main text as $\mathcal{C} = \left\{ \mathbf{C} \in \mathbb{R}^{N\times N} \middle| \, \text{var}(\mathbf{R}_0)\text{var}(\mathbf{f})\mathbf{C} = \mathbf{C}\text{var}(\mathbf{f})\text{var}(\mathbf{R}_0) \right\}$.

Proof. For any given matrix $\mathbf{C} \in \mathcal{C}$, I define the matrix

$$\mathbf{H} = (\mathbf{U}^{-1})^\top \mathbf{C} \mathbf{U}^{-1}. \tag{OA.16}$$

Then $\mathbf{H}$ satisfies

$$\mathbf{U}^\top \mathbf{U} \text{var}(\mathbf{f}) \mathbf{U}^\top \mathbf{H} \mathbf{U} = \mathbf{U}^\top \mathbf{H} \mathbf{U} \text{var}(\mathbf{f}) \mathbf{U}^\top \mathbf{U}, \tag{OA.17}$$

which simplifies to

$$\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top \mathbf{H} = \mathbf{H}\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top. \tag{OA.18}$$

I define the matrix

$$\mathbf{L} = \mathbf{G}^\top \mathbf{H} \mathbf{G}. \tag{OA.19}$$

Using equations (OA.1) and (OA.18), I have

$$\mathbf{L}\mathbf{\Pi} = \mathbf{G}^\top \mathbf{H}\mathbf{G}\mathbf{G}^\top \mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top \mathbf{G} = \mathbf{G}^\top \mathbf{H}\mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top \mathbf{G} = \mathbf{G}^\top \mathbf{U}\text{var}(\mathbf{f})\mathbf{U}^\top \mathbf{H}\mathbf{G} = \mathbf{\Pi}\mathbf{L}. \tag{OA.20}$$

For any vector $\mathbf{v}$ in the span of the j-th part of the partition of matrix $\mathbf{\Pi}$ in (OA.2), I have $\mathbf{\Pi}\mathbf{v} = \pi_j \mathbf{v}$. Therefore, I have

$$\mathbf{\Pi}(\mathbf{L}\mathbf{v}) = (\mathbf{\Pi}\mathbf{L})\mathbf{v} = (\mathbf{L}\mathbf{\Pi})\mathbf{v} = \mathbf{L}\pi_j \mathbf{v} = \pi_j(\mathbf{L}\mathbf{v}). \tag{OA.21}$$

Therefore, $\mathbf{L}\mathbf{v}$ is also a vector in the span of the j-th part of the partition. Because I can arbitrarily choose the vector $\mathbf{v}$ and the part $j = 1, 2, \ldots, J$, matrix $\mathbf{L}$ must be

$$\mathbf{L} = \begin{pmatrix} \mathbf{\Phi}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}_2 & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Phi}_J \end{pmatrix}, \tag{OA.22}$$

where each $\mathbf{\Phi}_j$ is an $r_j \times r_j$ matrix for $j = 1, 2, \ldots, J$.

The form (OA.22) and $\mathbf{L} = \mathbf{G}^\top \mathbf{H}\mathbf{G}$ imply that whenever the (l, l)-th and (m, m)-th elements of matrix $\mathbf{\Pi}$ in (OA.2) correspond to distinct eigenvalues, I have $\mathbf{g}_l^\top \mathbf{H}\mathbf{g}_m = 0$, where $\mathbf{g}_l$ and $\mathbf{g}_m$ are l-th and m-th column of $\mathbf{G}$. Recall the definitions $\mathbf{H} = (\mathbf{U}^{-1})^\top \mathbf{C}\mathbf{U}^{-1}$ and $\mathbf{B} = \mathbf{U}^{-1}\mathbf{G}$. Therefore, I have $\mathbf{b}_l^\top \mathbf{C}\mathbf{b}_m = 0$. □

The IUF condition therefore implies that

$$0 = \text{cov}(\mathbf{a}^\top \Delta\mathbf{p}(\mathbf{f}), s) = \text{cov}\left(\mathbf{b}_l^\top \text{var}(\mathbf{R}_0) \sum_{n=1}^{N} q_n \mathbf{h}_n, q_m\right) = \text{var}(q_m)\mathbf{b}_l^\top \text{var}(\mathbf{R}_0)\mathbf{h}_m. \tag{OA.23}$$

What remains follows the proof of Theorem 1 in the main text. Specifically, I project $\mathbf{h}_m$ onto the factor portfolios $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$ and obtain $\mathbf{h}_m = \sum_{n=1}^{N} \theta_{m,n} \mathbf{b}_n$. The IUF condition (OA.23) implies that $\theta_{m,n} = 0$, for any (n,n)-th and (m,m)-th elements of matrix $\mathbf{\Pi}$ that correspond to distinct eigenvalues. Therefore, I recover the block-diagonal form of the price-of-flow-induced-risk matrix of the F-SDF, as shown in (OA.4). Lastly, because of (OA.12) and the assumption of positive compensation for risk, $\mathbf{\Psi}_j$ is positive definite for all j. □

A.2 Optimality of the IUF Orthogonalization

This appendix shows the optimality of the IUF orthogonalization, without imposing regularity Assumption 4 in the main text.

In the general case, the definition of model orthogonalization in the main text needs to be modified to account for duplicate eigenvalues. The orthogonalized model now has the block-diagonal price-of-flow-induced-risk matrix that is consistent with (OA.2).

DEFINITION O.1. *Any N linearly independent portfolios $\tilde{\mathbf{B}} = (\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N)$ defines a model orthogonalization. The N^2 model expressed under portfolios $\tilde{\mathbf{B}}$ is*

$$\Delta \mathbf{p}(\mathbf{f}) = \text{var}(\mathbf{R}_0) \tilde{\mathbf{B}} \tilde{\mathbf{\Lambda}} \tilde{\mathbf{q}}, \tag{OA.24}$$

with portfolio flows $\tilde{\mathbf{q}} = (\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_N)^\top$ and the $N \times N$ price-of-flow-induced-risk matrix

$$\tilde{\mathbf{\Lambda}} = \begin{pmatrix} \tilde{\mathbf{\Psi}}_{1,1} & \tilde{\mathbf{\Psi}}_{1,2} & \cdots & \tilde{\mathbf{\Psi}}_{1,J} \\ \tilde{\mathbf{\Psi}}_{2,1} & \tilde{\mathbf{\Psi}}_{2,2} & \cdots & \tilde{\mathbf{\Psi}}_{2,J} \\ \cdots & \cdots & \cdots & \cdots \\ \tilde{\mathbf{\Psi}}_{J,1} & \tilde{\mathbf{\Psi}}_{J,2} & \cdots & \tilde{\mathbf{\Psi}}_{J,J} \end{pmatrix}, \tag{OA.25}$$

where each $\tilde{\mathbf{\Psi}}_{j,l}$ is an $r_j \times r_l$ matrix. The orthogonalized N-factor model under these portfolios is defined as

$$\Delta \bar{\mathbf{p}}(\mathbf{f}) = \text{var}(\mathbf{R}_0) \tilde{\mathbf{B}} \bar{\mathbf{\Lambda}} \tilde{\mathbf{q}}, \tag{OA.26}$$

with the $N \times N$ price-of-flow-induced-risk matrix

$$\bar{\mathbf{\Lambda}} = \begin{pmatrix} \tilde{\mathbf{\Psi}}_{1,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{\Psi}}_{2,2} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{\mathbf{\Psi}}_{J,J} \end{pmatrix}. \tag{OA.27}$$

I now generalize Theorem 2 of the main text.

Theorem O.2. *Fix any model orthogonalization $\tilde{\mathbf{B}} = (\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N)$. The orthogonalized model (OA.26) satisfies the IUF for any parameters $\tilde{\mathbf{\Psi}}_{1,1}, \tilde{\mathbf{\Psi}}_{2,2}, \ldots, \tilde{\mathbf{\Psi}}_{N,N}$ **if and only if** for any N^2 parameters $\tilde{\mathbf{\Lambda}}$ in (OA.25),*

- *a one-unit shock to portfolio flow $\tilde{q}_n$ causes the same amount of impact to the price of portfolio $\tilde{\mathbf{b}}_n$ under the N^2 model (OA.24) and the orthogonalized model (OA.26),*

$$\frac{\partial \tilde{\mathbf{b}}_n^\top \Delta \mathbf{p}(\mathbf{f})}{\partial \tilde{q}_n} = \frac{\partial \tilde{\mathbf{b}}_n^\top \Delta \bar{\mathbf{p}}(\mathbf{f})}{\partial \tilde{q}_n}. \tag{OA.28}$$

- *the arbitrageur's expected utility $\mathbb{E}[u(W(\mathbf{f}))]$ remains invariant under (OA.24) and (OA.26).*

Proof. The proof relies on three intermediate results.

The first result is that the orthogonalized model (OA.26) satisfies the IUF for any parameters $\tilde{\mathbf{\Psi}}_{1,1}, \tilde{\mathbf{\Psi}}_{2,2}, \ldots, \tilde{\mathbf{\Psi}}_{N,N}$ if and only if $\text{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ and $\text{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ whenever the (n,n)-th and (m,m)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues.

Proof. First, I show the sufficiency direction. From Theorem O.1, if the orthogonalized model (OA.26) satisfies the IUF, then there exists some $N \times N$ invertible matrix $\mathbf{O}$, such that for any $\bar{\mathbf{\Lambda}}$ in (OA.27), $\mathbf{O}^{-1}\bar{\mathbf{\Lambda}}\mathbf{O}$ is still a block-diagonal matrix of form (OA.27). I denote the (j,l)-th block of $\mathbf{O}^{-1}$ as $\mathbf{X}_{j,l}$, which is an $r_j \times r_l$ matrix. I denote the (j,l)-th block of $\mathbf{O}$ as $\mathbf{Y}_{j,l}$, which is an $r_j \times r_l$ matrix. I then know that $\mathbf{X}_{j,l}\tilde{\mathbf{\Psi}}_{l,l}\mathbf{Y}_{l,k} = \mathbf{0}$ for any l, any $j \neq k$, and

any parameter matrix $\tilde{\mathbf{\Psi}}_{l,l}$. This implies that at most one of the matrices $\mathbf{X}_{j,l}$ and $\mathbf{Y}_{l,k}$ is a non-zero matrix. Suppose that there exist some l and $k \neq k'$, such that both $\mathbf{Y}_{l,k}$ and $\mathbf{Y}_{l,k'}$ are non-zero matrices. Then all $\mathbf{X}_{j,l}$ are zero matrices for $j = 1, 2, \ldots, J$, which contradicts the fact that $\mathbf{O}$ is invertible. Therefore, for each l, at most one of the matrices from $\mathbf{Y}_{l,k}$ $(k = 1, 2, \ldots, J)$ is non-zero. Because $\mathbf{O}$ is invertible, for each k, at least one of the matrices from $\mathbf{Y}_{l,k}$ $(l = 1, 2, \ldots, J)$ is non-zero. Therefore, matrix $\mathbf{O}$ has exactly J non-zero blocks, which belong to distinct columns and rows. Moreover, all of these J blocks must be square matrices. Otherwise, I can find linearly dependent rows or columns of $\mathbf{O}$, contradicting the invertibility. Therefore, the matrix $\mathbf{O}$ is of the form

$$\mathbf{O} = \begin{pmatrix} \tilde{\mathbf{\Gamma}}_{1,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{\Gamma}}_{2,2} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{\mathbf{\Gamma}}_{J,J} \end{pmatrix} \tag{OA.29}$$

or its rearrangements, where each $\tilde{\mathbf{\Gamma}}_{j,j}$ is a $r_j \times r_j$ matrix. The only possible rearrangement of $\mathbf{O}$ is exchanging any two blocks of columns over for j and j', where the two rearranged blocks j and j' have the same dimension $r_j = r_{j'}$. Because $\tilde{\mathbf{B}}\mathbf{O} = \mathbf{B}$ and $\mathbf{O}\mathbf{q} = \tilde{\mathbf{q}}$, matrix $\mathbf{O}$ reorders blocks of portfolios $\tilde{\mathbf{B}} = (\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N)$ that correspond to distinct eigenvalues with the same degree of multiplicity and recombines portfolios that correspond to the same eigenvalue. The resulting portfolios $\mathbf{B}$ are the factor portfolios in Theorem O.1. Therefore, $\mathrm{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ and $\mathrm{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ whenever (n, n)-th and (m, m)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues.

Next, I show the necessity direction. Because $\mathrm{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ and $\mathrm{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ whenever the (n, n)-th and (m, m)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues, all I need for the form (OA.4) is to orthogonalize the portfolios $\tilde{\mathbf{B}} = (\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N)$ within each duplicate eigenvalue. I am free to do so, because there is no dimension reduction for the price-of-flow-induced-risk matrix within the same eigenvalue. The form (OA.4)

satisfies the IUF by Theorem O.1. □

The second claim is that (OA.28) holds for any $\tilde{\mathbf{\Lambda}}$ if and only if $\text{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ whenever (n,n)-th and (m,m)-th elements of $\mathbf{\Pi}$ correspond to distinct eigenvalues.

Proof. Note that I have by (OA.24),

$$\tilde{\mathbf{b}}_l^\top \Delta \mathbf{p}(\mathbf{f}) = \sum_{n=1}^{N} \sum_{m=1}^{N} \tilde{\lambda}_{n,m} \tilde{q}_m \text{cov}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0). \tag{OA.30}$$

I then have

$$\frac{\partial \tilde{\mathbf{b}}_l^\top \Delta \mathbf{p}(\mathbf{f})}{\partial \tilde{q}_l} = \sum_{n=1}^{N} \tilde{\lambda}_{n,l} \text{cov}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0). \tag{OA.31}$$

Similarly, I have by (OA.26),

$$\tilde{\mathbf{b}}_l^\top \Delta \bar{\mathbf{p}}(\mathbf{f}) = \sum_{n=1}^{N} \sum_{m=1}^{N} \bar{\lambda}_{n,m} \tilde{q}_m \text{cov}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0) \tag{OA.32}$$

and

$$\frac{\partial \tilde{\mathbf{b}}_l^\top \Delta \bar{\mathbf{p}}(\mathbf{f})}{\partial \tilde{q}_l} = \sum_{n=1}^{N} \bar{\lambda}_{n,l} \text{cov}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0). \tag{OA.33}$$

If the (n,n)-th and (l,l)-th elements of matrix $\mathbf{\Pi}$ correspond to the same eigenvalue, I have $\tilde{\lambda}_{n,l} = \bar{\lambda}_{n,l}$ by definition (OA.27). Therefore, (OA.31) and (OA.33) are equal for any $\tilde{\mathbf{\Lambda}}$ if and only if $\text{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ whenever (n,n)-th and (m,m)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues. □

The third claim is stated under the condition that $\text{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ whenever the (n,n)-th and (m,m)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues. The arbitrageur's expected utility $\mathbb{E}[u(W(\mathbf{f}))]$ remains invariant under models (OA.24) and (OA.26) if and only if $\text{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ whenever the (n,n)-th and (m,m)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues.

Proof. I partition the portfolio flows $\tilde{\mathbf{q}}$ according to eigenvalue values,

$$\tilde{\mathbf{q}}^\top = (\tilde{\mathbf{q}}_1, \tilde{\mathbf{q}}_2, \ldots, \tilde{\mathbf{q}}_J)^\top, \tag{OA.34}$$

where each $\tilde{\mathbf{q}}_j$ is an $r_j \times 1$ vector. Similarly, I partition the covariance $\tilde{\mathbf{B}}^\top \text{var}(\mathbf{R}_0)\tilde{\mathbf{B}}$ as

$$\tilde{\mathbf{B}}^\top \text{var}(\mathbf{R}_0)\tilde{\mathbf{B}} = \begin{pmatrix} \tilde{\mathbf{\Phi}}_{1,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{\Phi}}_{2,2} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{\mathbf{\Phi}}_{J,J} \end{pmatrix}, \tag{OA.35}$$

where each $\tilde{\mathbf{\Phi}}_{j,j}$ is an $r_j \times r_j$ matrix and I use $\text{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ whenever the (n,n)-th and (m,m)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues.

Under the price impact model (OA.24), the expected compensation is

$$\mathbb{E}[\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})] = \mathbb{E}[\tilde{\mathbf{q}}^\top \tilde{\mathbf{B}}^\top \text{var}(\mathbf{R}_0)\tilde{\mathbf{B}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{q}}] = \sum_{j=1}^{J}\sum_{l=1}^{J} \mathbb{E}\left[\tilde{\mathbf{q}}_j^\top \tilde{\mathbf{\Phi}}_{j,j}\tilde{\mathbf{\Psi}}_{j,l}\tilde{\mathbf{q}}_l\right]. \tag{OA.36}$$

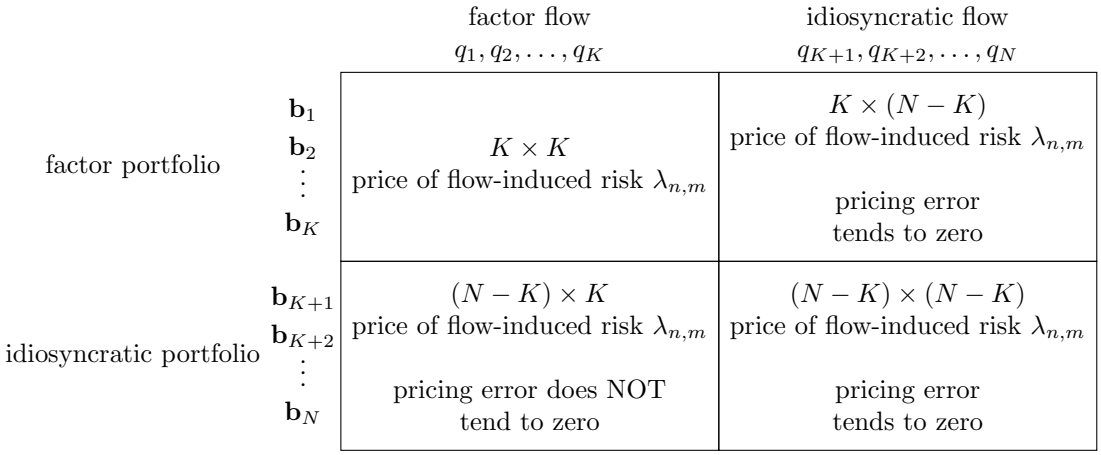
Similarly, under the price impact model (OA.26), the expected compensation is

$$\mathbb{E}[\mathbf{f}^\top \Delta\bar{\mathbf{p}}(\mathbf{f})] = \mathbb{E}[\tilde{\mathbf{q}}^\top \tilde{\mathbf{B}}^\top \text{var}(\mathbf{R}_0)\tilde{\mathbf{B}}\bar{\mathbf{\Lambda}}\tilde{\mathbf{q}}] = \sum_{j=1}^{J} \mathbb{E}\left[\tilde{\mathbf{q}}_j^\top \tilde{\mathbf{\Phi}}_{j,j}\tilde{\mathbf{\Psi}}_{j,j}\tilde{\mathbf{q}}_j\right]. \tag{OA.37}$$

As shown in the main text, the arbitrageur's expected utility $\mathbb{E}[u(W(\mathbf{f}))]$ remains invariant under models (OA.24) and (OA.26) if and only if the expected compensation $\mathbb{E}[\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})]$ remains invariant. Because $\tilde{\mathbf{\Psi}}_{j,l}$ are free parameters, the expected compensation (OA.36) and (OA.37) are the same if and only if $\text{cov}(\tilde{\mathbf{q}}_j, \tilde{\mathbf{q}}_l) = \mathbf{0}$ for any $j \neq l$. The last statement is precisely $\text{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ whenever the (n,n)-th and (m,m)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues. □

The three claims together imply that Theorem O.2 is true. □

Figure O.1. Necessity of the IUF to the flow-based APT



Notes: This figure shows why the IUF is necessary for the flow-based APT. Without the IUF, the price impact of factor flows on idiosyncratic portfolios does not tend to zero, even when the variance of idiosyncratic flows tends to zero. This effect is illustrated in the bottom-left block of the price-of-flow-induced-risk matrix.

B Necessity of the IUF to the Flow-Based APT

One may wonder if there exists a version of the flow-based APT that does not impose the IUF and directly reduces the N^2 price impact model to some K^2 model. That is, I start with the N^2 model

$$\Delta \mathbf{p}(\mathbf{f}) = \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_{n,m} q_m \text{cov}(\mathbf{R}_0, \mathbf{b}_n^\top \mathbf{R}_0), \tag{OB.1}$$

and aim to approximate this model by the first K factors,

$$\Delta \check{\mathbf{p}}(\mathbf{f}) = \sum_{k=1}^{K} \sum_{j=1}^{K} \lambda_{k,j} q_j \text{cov}(\mathbf{R}_0, \mathbf{b}_k^\top \mathbf{R}_0). \tag{OB.2}$$

Can I bound the pricing error $\Delta \mathbf{p}(\mathbf{f}) - \Delta \check{\mathbf{p}}(\mathbf{f})$ in a similar manner to the flow-based APT in the main text?

The answer is no. Figure O.1 illustrates the situation. The bound in the main text on the volatility of the F-SDF implies a uniform upper bound on the N^2 prices of flow-induced risk $\lambda_{n,m}$. The flow-based APT assumes the variance $\sum_{n=K+1}^{N} \pi_n$ of idiosyncratic flows $q_{K+1}, q_{K+2}, \ldots, q_N$ tending to zero. This assumption ensures that the pricing error

caused by idiosyncratic flows, which is $\sum_{n=1}^{N}\sum_{m=K+1}^{N}\lambda_{n,m}q_m\text{cov}(\mathbf{R}_0,\mathbf{b}_n^\top\mathbf{R}_0)$, tends to zero, which corresponds to the top-right and bottom-right blocks of Figure O.1.

However, note that

$$\Delta\mathbf{p}(\mathbf{f})-\Delta\check{\mathbf{p}}(\mathbf{f})=\sum_{n=1}^{N}\sum_{m=K+1}^{N}\lambda_{n,m}q_m\text{cov}(\mathbf{R}_0,\mathbf{b}_n^\top\mathbf{R}_0)+\sum_{n=K+1}^{N}\sum_{m=1}^{K}\lambda_{n,m}q_m\text{cov}(\mathbf{R}_0,\mathbf{b}_n^\top\mathbf{R}_0). \tag{OB.3}$$

The second term is the price impact of factor flows on portfolios that idiosyncratic flows go into, which does not tend to zero, even when the variance of idiosyncratic flows tends to zero. This component of pricing error corresponds to the bottom-left block of Figure O.1. While flows $q_{K+1},q_{K+2},\ldots,q_N$ are idiosyncratic, portfolios $\mathbf{b}_{K+1},\mathbf{b}_{K+2},\ldots,\mathbf{b}_N$ that correspond to these flows may be important risk factors. Factor flows $q_1,q_2,\ldots,q_K$ could have a large cross-impact on these portfolios. Without the IUF, one lacks a proper theoretical justification for eliminating this bottom-left component.

Instead, my IUF approach first chooses the specific flows $q_1,q_2,\ldots,q_N$ and corresponding portfolios $\mathbf{b}_1,\mathbf{b}_2,\ldots,\mathbf{b}_N$ that have both uncorrelated flows and uncorrelated fundamental returns. The IUF implies that these portfolios have no cross-impacts. That is, the IUF first reduces the N^2 prices of flow-induced risk to the N diagonal terms. The flow-based APT then reduces these N prices of flow-induced risk to K using the factor structure of flows.