FTG Working Paper Series

Flow-Based Arbitrage Pricing Theory

by

Yu An

Working Paper No. 00099-00

Finance Theory Group

www.financetheory.com

Flow-Based Arbitrage Pricing Theory[*]

Yu An

This version: April 28, 2023
First version: May 2, 2022

Abstract

I introduce a new approach, model, and definition for analyzing demand effects in asset pricing. My approach generalizes arbitrage pricing, and avoids making any parametric assumptions on utility function and payoff distribution, which are commonly found in equilibrium literature. My approach also reveals and relaxes an unrealistic cross-sectional restriction on price impacts in the literature's quadratic-normal setup. In my model, price impacts between underlying assets occur through factors, which connect to these assets through the covariance structure of noisy flows and fundamental risks. Specifically, I develop a new definition for factor-level demand elasticity, highlighting that the conventional definition is ill-defined.

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

JEL Codes: G11, G12

[*] Carey Business School, Johns Hopkins University; yua@jhu.edu. I especially thank Yinan Su and Chen Wang for numerous conversations. I thank Yun Joo An, Daniel Andrei, Federico Bandi, Darrell Duffie, Thierry Foucault, Nicola Fusari, Ben Hébert, Steve Heston, Arvind Krishnamurthy, Praveen Kumar, Mina Lee, Jane Li, Jiacui Li, Wenhao Li, Daniel Neuhann, Jun Pan, Nagpurnanand Prabhala, Matthew Pritsker, Marzena Rostek, Alessandro Rebucci, Shrihari Santosh, Dongho Song, Zhaogang Song, Semih Üslü, Adrien Verdelhan, Milena Wittwer, Tony Zhang, Zeyu Zheng, and seminar participants at JHU Carey economics and finance brownbag, USC finance reading group, Australasian finance & banking conference, MFA annual conference, University of Colorado Boulder, Eastern Finance Association annual conference, and FMCG conference for helpful comments. All errors are my own.

1 Introduction

In economics, demand and supply determine price and quantity. In finance, the absence of arbitrage ensures the existence of stochastic discount factors (SDF). However, arbitrage pricing is long believed to be incompatible with demand effects, because the arbitrage theory developed since Black and Scholes (1973) and Merton (1973b) models only price, but not quantity. The failure of arbitrage in the presence of noise traders is often referred to as "the limits of arbitrage," suggesting that arbitrageurs cannot conduct textbook arbitrage when facing frictions that limit their ability to absorb flows (De Long, Shleifer, Summers, and Waldmann, 1990; Shleifer and Vishny, 1997; Du, Tepper, and Verdelhan, 2018).

This paper has three logically connected contributions. First, since 1980s, the literature has used equilibrium approach to model demand effects, particularly using the quadratic-normal setup.[1] Instead of this approach, I generalize arbitrage pricing to model demand effects, and avoid making any parametric assumption regarding the utility function and payoff distribution. Crucially, by abstracting away from specific economic environments, I can characterize the general price/quantity relationship across multiple risky assets. My analysis reveals that, within the quadratic-normal setup, arbitrageurs exhibit equal aversion to risks induced by flows into different portfolios. By relaxing this restrictive assumption, I develop a new and more flexible model.

Second, the literature uses price elasticity or cross elasticity to study how the flow into asset A impacts the price of asset A or B (Koijen and Yogo, 2019; Gabaix and Koijen, 2022). In my model, the flow into asset A impacts the price of asset B through factors, which are generally long-short portfolios of underlying assets and are connected to these assets via the covariance structure of noisy flows and fundamental risks. Contrary to the asset-by-asset isolated approach, my factor model integrates demand effects with cross-sectional asset

[1] 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.

pricing—individual assets' price impacts and cross impacts are connected through factors.

Third, the literature uses $(dP/P)/(dQ/Q)$ to define elasticity for factors, measuring how a 1% change in quantity affects price.[2] While this definition is suitable for individual assets, I show that it is mathematically ill-defined and economically nonsensical *when applied to factors (i.e., long-short portfolios).* In my model, the definition of factor-level elasticity involves explicitly constructing risk exposures, and measures how one unit of risk induced by the quantity change affects price. This distinction is not merely a technical point, but reflects this paper's core economics—price impacts arise because flow-absorbing arbitrageurs are averse to risk, not because arbitrageurs are inelastic to flow per se.[3]

In what follows, I present five new tools to derive my factor model of price impacts.

First, I introduce the flow-based SDF. Drawing from the framework presented in Cochrane (2009) textbook, I examine N assets with random payoffs $\mathbf{X}$ at time 1, and analyze asset prices $\mathbf{P}$ at time 0. The law of one price (LOOP) implies the existence of some SDF M, which is a time-1 random variable, such that $\mathbf{P} = \mathbb{E}[M\mathbf{X}]$.

I diverge from Cochrane (2009) by assuming the presence of noisy flows[4] $\mathbf{f} = (f_1, f_2, \ldots, f_N)^\top$ into these assets at time 0. By "noisy," I refer to the fact that payoffs $\mathbf{X}$ are independent of flows $\mathbf{f}$, and thus, the flows influence time-0 prices $\mathbf{P}$ solely through demand effects. Arbitrage pricing does not model specific economic agents; however, one can envision a representative arbitrageur who faces certain "limits of arbitrage" and accommodates $\mathbf{f}$. Consequently, asset prices $\mathbf{P}(\mathbf{f})$ and the SDF $M(\mathbf{f})$ may now depend on $\mathbf{f}$.

By comparing the time-0 economy with flow $\mathbf{f}$ to the one without any flow, I obtain

$$\underbrace{\Delta\mathbf{p}(\mathbf{f})}_{\text{price impact}} = \mathbb{E}[\underbrace{\tilde{M}(\mathbf{f})}_{\text{F-SDF}} \underbrace{\mathbf{R}_0}_{\text{fundamental return}}], \tag{1}$$

[2] For example, this definition is used by Table 1 of Gabaix and Koijen (2022), which also provides a summary of estimates in the empirical literature.

[3] That risk matters for price impacts is empirically documented by, for example, Wurgler and Zhuravskaya (2002) and Li, Fu, and Chaudhary (2022) in stock and corporate bond markets.

[4] Bold font is used to represent matrices and vectors, with $\mathbf{b}^\top$ denoting the transpose of $\mathbf{b}$.

where $\Delta\mathbf{p}(\mathbf{f}) := (\mathbf{P}(\mathbf{f}) - \mathbf{P}(\mathbf{0}))/\mathbf{P}(\mathbf{0})$ (element-wise division) represents the assets' time-0 price impacts, and $\mathbf{R}_0 := \mathbf{X}/\mathbf{P}(\mathbf{0})$ represents the assets' fundamental-driven returns between time 0 and 1. Notably, the *flow-based SDF (F-SDF)*, defined as $\tilde{M}(\mathbf{f}) := M(\mathbf{f}) - M(\mathbf{0})$, describes how the flow changes the standard SDF.

The insight is that to characterize 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 $\tilde{M}(\mathbf{f}) = M(\mathbf{f}) - M(\mathbf{0})$, rather than tackling the more complex question of how the entire SDF $M(\mathbf{f})$ depends on $\mathbf{f}$. Economically, $\tilde{M}(\mathbf{f})$ encodes all the risk preferences of how the arbitrageur responds to $\mathbf{f}$.

Second, I develop the portfolio theory for flows. The basis of asset pricing is portfolio theory (Markowitz, 1952). Surprisingly, a theory for constructing portfolio flows does not exist in the literature. Empirical research has employed ad hoc constructions of portfolio flows, similar to the approach taken for portfolio returns. I show that the portfolio flows are equal to the *pseudoinverse* of portfolio weights times asset flows. Intuitively, the economic interpretation of portfolio flows is *flows into portfolios*—portfolio flows first enter the portfolios and are then allocated to individual assets according to portfolio weights. Consequently, one should project asset flows onto portfolio weights to obtain portfolio flows, which explains the use of pseudoinverse construction. On a deeper level, the total quantity of risk absorbed by arbitrageurs, represented by the inner product $\mathbf{f}^\top\mathbf{R}_0$ between flows and returns, should remain invariant under portfolio formation. I prove that this invariance principle implies that the construction of portfolio flows should follow my proposed procedure, which contrasts with the construction of portfolio returns.

Third, I present the flow-based Hansen-Jagannathan bound. A common intuition is that "noisy flows shock the SDF and change asset prices today, leading to flow-induced return predictability in the future." My flow-based bound formalizes this intuition. I define any portfolio's price impact ratio as the portfolio's price impact over the portfolio's fundamental risk. I prove that the maximum price impact ratio (MPIR) over all portfolios equals the minimum volatility of the F-SDF. Intuitively, shorting the MPIR portfolio is the mean-variance optimal strategy to provide liquidity to flows. My result connects this optimal portfolio of liquidity providers to the volatility of flow-induced changes in the standard SDF.

Fourth, I develop a new cross-sectional restriction for price impacts. Intuitively, not all price impacts in the cross-section are plausible. For a simple example, consider two assets with uncorrelated flows $\mathbf{f}$ and uncorrelated fundamental returns $\mathbf{R}_0$. In this case, one should not expect a flow into the first asset to impact the second asset's price. *Irrelevance of Uncorrelated Flows (IUF)* formalizes this intuition by eliminating all such implausible price impacts in a setting with general covariance structures of flows and fundamental returns.

I prove that the IUF holds if and only if the variations of the N assets' price impacts are driven by N orthogonalized factors that have no cross-impacts between each other. These factors are endogenously and (generally) uniquely chosen to have both uncorrelated flows and uncorrelated fundamental returns. Each factor's price elasticity to its flow is characterized by a free parameter.

Mathematically, I achieve the IUF characterization by employing commuting matrices. Roughly speaking, two square matrices $\mathbf{A}$ and $\mathbf{B}$ commute (i.e., $\mathbf{AB} = \mathbf{BA}$) if and only if they share the same eigenvectors. Applying commuting matrices to the covariance structures of noisy flows and fundamental returns enables me to uniquely pin down asset-pricing factors (i.e., eigenvectors). To the best of my knowledge, commuting matrices have not been previously applied in economics or finance.

A natural question arises as to why the IUF eliminates cross-impacts between the specific factors that exhibit both uncorrelated flows and uncorrelated fundamental returns. This selection of factors is not only intuitive but also optimal. I prove that, among all possible ways of choosing N factors to span the N assets' price impacts, the IUF is the only approach that preserves each factor's price elasticity as invariant and preserves the arbitrageur's expected utility as invariant for any specific utility function.

Fifth, I develop the flow-based Arbitrage Pricing Theory (APT) to reduce the N factors selected by the IUF to K systematic factors. I assume that *(i)* flows exhibit a low-dimensional factor structure and *(ii)* small flows into a portfolio with little fundamental risks should not have a high price impact. Condition *(i)* is well supported empirically.[5] Condition *(ii)* forms the core economics underlying my arbitrage theory with flows, which is characterized using the price impact ratio and the flow-based Hansen-Jagannathan bound. This condition generalizes the "good-deal bound" of Ross (1976) and Cochrane and Saa-Requejo (2000). I prove that under these two conditions, the K-factor model approximates the N-factor model in the mean-squared-error sense.

The five new theoretical tools allow me to derive the following new model, which, intuitively, combines the canonical factor model and price elasticity,

$$\underbrace{\mathbf{f}}_{\text{asset flow}} = \sum_{k=1}^{K} \underbrace{\mathbf{b}_k}_{\text{flow beta}} \underbrace{q_k}_{\text{factor flow}} + \underbrace{\mathbf{e}}_{\text{idiosyncratic flow}}, \tag{2}$$

$$\underbrace{\Delta\mathbf{p}(\mathbf{f})}_{\text{price impact}} = \sum_{k=1}^{K} \underbrace{\lambda_k}_{\text{price of flow-induced risk}} \underbrace{q_k}_{\text{factor flow}} \underbrace{\operatorname{cov}(\mathbf{R}_0, \mathbf{b}_k^\top \mathbf{R}_0)}_{\text{quantity of risk}}. \tag{3}$$

Equation (2) decomposes the N-asset flows $\mathbf{f} = (f_1, f_2, \ldots, f_N)^\top$ into K-factor flows $q_1, q_2, \ldots, q_K$ and idiosyncratic flows $\mathbf{e} = (e_1, e_2, \ldots, e_N)^\top$. The new insight is that *flow betas are portfolio weights*. Note that if factor-k flow q_k increases by \$1, asset-n flow increases by \$$b_{n,k}$. By market clearing, a one-dollar factor-k flow decreases the arbitrageurs' holding in asset n by \$$b_{n,k}$. Thus, flow betas $\mathbf{b}_k = (b_{1,k}, b_{2,k}, \ldots, b_{N,k})^\top$ mimic the changes in the arbitrageurs' portfolio holding caused by one dollar of factor-k flow, implying that $\mathbf{b}_k^\top \mathbf{R}_0$ is the risk factor corresponding to the factor flow q_k. This result is connected to and generalizes the approach of Alekseev, Giglio, Maingi, Selgrad, and Stroebel (2022), who use the beta of mutual fund portfolio changes in reaction to heat shocks to construct hedging portfolios for climate risk.

Equation (3) implies that the factor flows q_k drive the cross-section of price impacts $\Delta\mathbf{p}(\mathbf{f})$. The quantity of risk exposure of individual assets to these factors is the return covariance $\operatorname{cov}(\mathbf{R}_0, \mathbf{b}_k^\top \mathbf{R}_0)$. The new parameter $\lambda_k > 0$, termed *the price of flow-induced*

[5] Hasbrouck and Seppi (2001) first document the factor structure of flows. Their finding is motivated using optimization models of trading (Caballe and Krishnan, 1994; Kumar and Seppi, 1994).

risk, measures for how one unit of risk induced by the flow changes the price of factor k.

In comparison to the existing literature, my model is developed under less restrictive conditions, dispensing with the need for parametric assumptions on the utility function and payoff distribution. Moreover, as outlined below, my model provides five key improvements.

First, my model relaxes the literature's standard model. I prove that the static quadratic-normal model is a special case of my model (3) when $K = N$ and the price of flow-induced risk is equal for all factors, i.e., $\lambda_1 = \lambda_2 = \cdots = \lambda_K = \gamma/(\mu R_F)$. Here, γ is the arbitrageurs' CARA risk aversion, μ is the mass of arbitrageurs, and R_F is the gross risk-free rate.

$$\begin{aligned}&\text{my factor model} = \text{IUF} + \text{flow-based APT},\\&\text{static quadratic-normal model} = \text{IUF} + \text{equal-}\lambda_k.\end{aligned}$$

The assumption of equal-λ_k is both restrictive and unrealistic. Theoretically, equal-λ_k implies that the ratio of risk aversion to the mass of arbitrageurs must be identical across different factors. Empirically, An, Su, and Wang (2022) estimate the Fama and French (1993) three factors' λ_k using U.S. equity mutual fund flow and reject the equal-λ_k restriction.

The static quadratic-normal model has always implicitly assumed the IUF, although this underlying assumption goes unnoticed in the equilibrium literature. My approach brings to light the previously obscured cross-sectional relationship between flow and risk. Economically, the static quadratic-normal model combines a reasonable assumption (IUF) with a less tenable one (equal-λ_k). As such, I propose a more plausible and empirically grounded alternative (flow-based APT) to replace the latter assumption.

Second, my model serves as a natural dynamic extension of the static quadratic-normal model. Although the current setting is static, my model could be microfounded in a dynamic framework incorporating the predictability of future flows. Specifically, An and Zheng (2023) examine a standard dynamic consumption problem in which a representative agent also absorbs noisy flows into different assets. They show that the predictability of today's flow for tomorrow's flow permits differential λ_k for distinct factors, thereby eliminating the equal-λ_k constraint embedded in the static model. Analogously, my paper generalizes the Ross (1976) APT approach to microfound the factor model (3), while An and Zheng (2023) generalize the Merton (1973a) ICAPM approach.

Third, my model uncovers a novel cross-asset price impact channel.[6] I illustrate this channel through a commonality-in-flow paradox. Consider two assets, A and B, with uncorrelated fundamental returns but correlated flows. My model implies that, when controlling for the flow into asset A, the flow into asset B still influences the price of asset A. The static quadratic-normal model does not predict such cross-impact. In this paradox, controlling for asset A's flow is essential, as it eliminates the mechanical effect where asset B's flow alters asset A's flow, subsequently affecting asset A's price.

This new channel emerges precisely because my model has more degrees of freedom in λ_k than the static quadratic-normal model. In my model, asset B's flow changes the prices of different factors, which in turn change asset A's price by no arbitrage. In the special quadratic-normal setup where all factors have identical λ_k, the changing prices of different factors precisely counterbalance each other, leaving asset A's price unaltered.

Fourth, my model correctly defines factor-level price elasticity, addressing the shortcomings of the traditional approach. The literature measures price elasticity using the formula $(dP/P)/(dQ/Q)$, which represents how a 1% change in quantity impacts the price. The issue with this formula is that it is mathematically ill-defined and economically nonsensical for factors, which are generally long-short portfolios. Since the total quantity Q of a long-short portfolio is zero,[7] the denominator dQ/Q is ill-defined, as one cannot divide by zero.[8]

[6] For examples of cross-impact models in the literature, see Kodres and Pritsker (2002), Basak and Pavlova (2013), Pasquariello and Vega (2015), and Buffa and Hodor (2023).

[7] The total quantity Q of a long-short portfolio is calculated in a manner similar to portfolio flow dQ. This involves projecting the asset-level amount outstanding onto a set of portfolio weights. In most empirical exercises, the market factor is included, and its portfolio weights are perfectly collinear with the asset-level amount outstanding. As a result, the total quantity Q of the market factor is the sum of all assets, as typically expected. However, for all other portfolios, the total quantity Q is zero.

[8] The empirical literature tries to circumvent the issue of $Q = 0$ by constructing portfolio flow as $dQ/Q := \sum_n w_n dS_n/S_n$, where w_n is the portfolio weight, and dS_n/S_n is the asset-level flow in percentages. However, as previously mentioned, this construction is flawed because it mistakenly treats flows as returns.

In contrast, my definition $\lambda = (dP/P)/(dQ\text{var}(R))$ measures the impact of one unit of risk induced by the quantity change on the factor price, rather than the impact of a 1% change in quantity. The improvements over the traditional definition include: *(i)* I establish a theory that correctly constructs portfolio flow dQ, and *(ii)* I incorporate risk exposure $\text{var}(R)$, because the economic channel is that flow alters risk exposure, which subsequently leads to changes in factor price.

Fifth, my model holds direct empirical relevance. A large strand of empirical literature constructs proxies for noisy flows and investigates their pricing implications.[9] Derived from first principles, my model offers a theoretical foundation for a new set of regressions that estimate the price/quantity relationship for the cross-section of assets via factors. Just like my model generalizes the canonical factor model, the new regressions, implemented in An, Su, and Wang (2022), generalize the canonical Fama and MacBeth (1973) regression.

My factor model (3) imposes an empirical restriction—price impacts that arise from arbitrageurs' risk aversion do not depend on idiosyncratic flows (represented by the residual $\mathbf{e}$ in equation (2)). Intuitively, arbitrageurs absorb flows into all assets simultaneously, rather than considering each asset in isolation. Consequently, from the arbitrageurs' portfolio perspective, systematic flows are significant while idiosyncratic flows are not. This intuition generalizes the canonical understanding of risks, which posits that expected returns do not depend on idiosyncratic risks.

An, Su, and Wang (2022) derive a direct empirical test and validate this prediction using mutual fund flow data.[10] They prove that the χ^2 test statistic for the price impacts of idiosyncratic flows is associated with the corresponding maximum price impact ratios (MPIR). This test is analogous to and generalizes the canonical Gibbons, Ross, and Shanken (1989) test, which relates the anomaly expected returns of idiosyncratic risks to the corresponding maximum Sharpe ratios.

[9] See, for example, Coval and Stafford (2007), Frazzini and Lamont (2008), and Lou (2012).

[10] This prediction is also empirically supported by Cremers and Mei (2007) and Li and Lin (2022), who use alternative empirical tests and data.

My theory also offers a straightforward approach to quantify the contribution of flows to the volatility of the SDF. By the flow-based Hansen-Jagannathan bound, this contribution is lower bounded by the MPIR in the data. An, Su, and Wang (2022) estimate an annualized MPIR of approximately 0.5, suggesting that mutual fund flows can account for a significant portion of the SDF variation.

2 Related Literature

Asset pricing concerns risk, which is understood through portfolios (Markowitz, 1952). Yet, the demand-based asset pricing literature lacks a portfolio theory that accounts for both price and quantity. This gap in theory means that the empirical literature, which relies on heuristic portfolio construction to study demand effects, may not be accurate. Specifically, the literature either neglects or provides incorrect answers to three fundamental questions: *(i)* how to construct portfolio flows? *(ii)* how to define price elasticity for long-short portfolios? *(iii)* what implicit restrictions does the quadratic-normal setup impose on different portfolios' price elasticity? I answer all three, establishing a rigorous theoretical foundation that bridges cross-sectional asset pricing and demand-based asset pricing.

My approach integrates noise trading with factors, drawing inspiration from Kozak, Nagel, and Santosh (2018). While their model offers valuable insights, it suffers from the equal-λ_k issue that is prevalent in other models in the literature. To address this limitation, my model relaxes the equal-λ_k constraint, which represents more than just a technical contribution. By relaxing this restriction, the quadratic-normal model of price impacts transforms from being a microstructure model to one that is capable of generalizing the major asset-pricing results outlined in Cochrane (2009). This increased flexibility of λ_k enables empirical estimation of price impacts for the cross-section of assets using factors.

The arbitrage theory, which has developed since the works of Black and Scholes (1973), Merton (1973b), and Ross (1976), assumes perfect arbitrage and focuses solely on price modeling. My contribution to this field is generalizing arbitrage theory to model both price and quantity, thereby accounting for more realistic demand effects and the concept of "limits of arbitrage." By jointly modeling the factor structures of price and quantity, my theory distinguishes itself from existing generalizations of APT (e.g., Chamberlain, 1983, Cochrane and Saa-Requejo, 2000, and Raponi, Uppal, and Zaffaroni, 2022).

My arbitrage approach to demand effects complements the equilibrium approach (Koijen and Yogo, 2019; Gabaix and Koijen, 2022). Equilibrium models are typically solved in two steps. First, the response of a given arbitrageur to the trading flows of others is determined by assuming a specific utility function and payoff distribution, commonly quadratic-normal. Second, the actions of different arbitrageurs are aggregated to establish the equilibrium price and quantity. By generically characterizing the relationship between price and quantity, my arbitrage approach focuses on and enhances the first step of this process, which serves as the foundation for any competitive equilibrium model that incorporates demand effects.

The term "flow-based SDF" was introduced in section 5.3 of Gabaix and Koijen (2022). They solve a general equilibrium model and express the model solutions using an SDF that depends on flows. Additionally, many papers simply assume certain functional forms of how flows enter the SDF. In contrast, I *derive* the way in which flows enter the SDF by employing a set of axiomatic assumptions on price impacts. No existing paper arrives at my F-SDF pricing equation (1), which serves as the starting point for all my analyses.

In the microstructure literature, a technical trick is involves rotating assets to portfolios with uncorrelated fundamental risks and assuming that signals on these uncorrelated risks are also uncorrelated (see chapter 3.8 of Veldkamp, 2011). My marginal contributions can be found in two key areas. First, I develop a portfolio theory that correctly rotates trading flows. Second, by rotating to portfolios with both uncorrelated fundamental risks and uncorrelated flows, I determine how arbitrageurs respond to flow-induced risk in the cross-section through differential λ_k. This contrasts with the microstructure literature's quadratic-normal setup, which implicitly assumes equal λ_k for all portfolios.

In my model, price impacts emerge due to arbitrageurs' aversion to absorbing risk, rather than being driven by microstructure or liquidity frictions. This feature distinguishes my model from the existing literature that investigates the effects of commonality in liquidity on expected returns (Chordia, Roll, and Subrahmanyam, 2000; Hasbrouck and Seppi, 2001; Pástor and Stambaugh, 2003).

Dou, Kogan, and Wu (2021) and Kim (2020) show that the commonality in flows in and out of mutual funds is a priced risk factor. Lo and Wang (2000) establish the factor structure of trading volume, and Alvarez and Atkeson (2018) consider when volume becomes a risk factor. In contrast to these studies, my model posits that common flows into the cross-section of assets do not constitute risk factors but rather impact prices through demand effects.

3 Flow-Based Arbitrage Pricing Theory

In this section, I present flow-based arbitrage pricing theory.

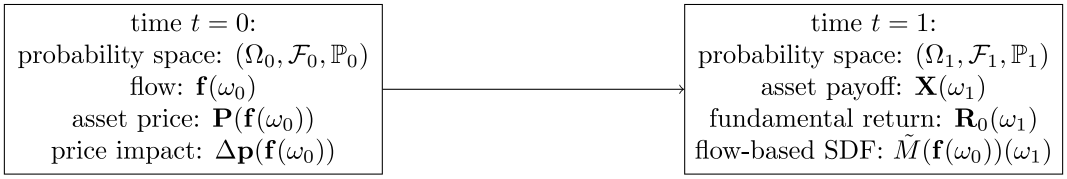
3.1 Flow-Based SDF

Model setup. My model consists of two periods, $t = 0$ and $t = 1$. The probability space is $(\Omega, \mathcal{F}, \mathbb{P}) = (\Omega_0 \times \Omega_1, \mathcal{F}_0 \times \mathcal{F}_1, \mathbb{P}_0 \times \mathbb{P}_1)$. I use ω_0 and ω_1 to denote elements of Ω_0 and Ω_1, respectively. To avoid ambiguity, I explicitly specify the dependence of random variables on ω_0 and ω_1 when necessary.

The model includes N assets. The flow into asset $n = 1, \dots, N$ is represented by $f_n(\omega_0)$, a random variable that realizes at time 0, with $\mathbf{f} = (f_1, f_2, \dots, f_N)^\top$. The unit of flow is one dollar. The payoff of asset n is denoted by $X_n(\omega_1)$, a random variable that realizes at time 1, with $\mathbf{X} = (X_1, X_2, \dots, X_N)^\top$.The flow is noisy in the sense that $\mathbf{f}$ is independent of the payoff $\mathbf{X}$. In other words, the flow influences asset prices solely through demand effects, and I do not model information effects. The gross risk-free rate is a constant, R_F, and does not depend on the flow.

I denote the time-0 price of asset n as $P_n(\mathbf{f}(\omega_0))$, with $\mathbf{P}(\mathbf{f}) = (P_1(\mathbf{f}), P_2(\mathbf{f}), \dots, P_N(\mathbf{f}))^\top$.

Figure 1. Model timeline


time $t = 0$:
probability space: $(\Omega_0, \mathcal{F}_0, \mathbb{P}_0)$
flow: $\mathbf{f}(\omega_0)$
asset price: $\mathbf{P}(\mathbf{f}(\omega_0))$
price impact: $\Delta\mathbf{p}(\mathbf{f}(\omega_0))$
time $t = 1$:
probability space: $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$
asset payoff: $\mathbf{X}(\omega_1)$
fundamental return: $\mathbf{R}_0(\omega_1)$
flow-based SDF: $\tilde{M}(\mathbf{f}(\omega_0))(\omega_1)$

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 realization of flow $\mathbf{f}(\omega_0)$, the flow-based SDF $\tilde{M}(\mathbf{f}(\omega_0))$ is a random variable that realizes at time 1.

In contrast to the equilibrium approach, the arbitrage approach does not model specific economic agents. However, one can envision a representative arbitrageur who absorbs the flow $\mathbf{f}$, allowing the price $\mathbf{P}(\mathbf{f})$ to depend on $\mathbf{f}$. Figure 1 displays the model timeline.

Without loss of generality, I assume that $\text{var}(\mathbf{X})$ and $\text{var}(\mathbf{f})$ have full rank.[11] I also assume that $\mathbb{E}[\mathbf{f}] = \mathbf{0}$. My theory focuses solely on the unexpected flow.[12] Importantly, I do not impose any specific distributional assumptions on flow or payoff.

The novelty of my framework lies in the fact that different realizations of flow $\mathbf{f}(\omega_0)$ correspond to distinct states of the world ω_0 at time 0. This setup is not only tractable but also possesses a strong economic rationale. In practice, short-term price impacts generated by noisy flows typically take a long time to revert before fundamental payoffs realize. In my model, the transition from time 0 to time 1 represents this lengthy period during which asset payoffs are largely independent of flows. Endowing flows with a distribution at time 0, rather than inserting an extra time period, highlights the distinction between short-term and long-term economic considerations.

Flow-based SDF. The LOOP implies that for any given flow $\mathbf{f}$, there exists a standard

[11] Otherwise, one can select linearly independent portfolios and rotate the payoff $\mathbf{X}$, flow $\mathbf{f}$, and price $\mathbf{P}(\mathbf{f})$ to those portfolios. See Section 3.2 for details on constructing portfolio flows.

[12] For the price impacts of expected flow, see Vayanos (2021) and Hartzmark and Solomon (2022).

SDF $M(\mathbf{f})$, which is a random variable that realizes at time 1, such that[13]

$$\mathbf{P}(\mathbf{f}) = \mathbb{E}[M(\mathbf{f})\mathbf{X}]. \tag{4}$$

It is important to 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

$$\mathbf{P}(\mathbf{0}) = \mathbb{E}[M(\mathbf{0})\mathbf{X}]. \tag{5}$$

For any flow $\mathbf{f}(\omega_0)$, I define *price impact* as

$$\Delta\mathbf{p}(\mathbf{f}(\omega_0)) := \left(\frac{P_1(\mathbf{f}(\omega_0)) - P_1(\mathbf{0})}{P_1(\mathbf{0})}, \frac{P_2(\mathbf{f}(\omega_0)) - P_2(\mathbf{0})}{P_2(\mathbf{0})}, \dots, \frac{P_N(\mathbf{f}(\omega_0)) - P_N(\mathbf{0})}{P_N(\mathbf{0})}\right)^\top, \tag{6}$$

which represents the percentage price change at time 0 with and without flow $\mathbf{f}(\omega_0)$. I explicitly write out the dependence on ω_0 to emphasize the fact that the price impact $\Delta\mathbf{p}(\mathbf{f})$ is a random vector that realizes at time 0. I define *fundamental return* as

$$\mathbf{R}_0 := \left(\frac{X_1}{P_1(\mathbf{0})}, \frac{X_2}{P_2(\mathbf{0})}, \dots, \frac{X_N}{P_N(\mathbf{0})}\right)^\top, \tag{7}$$

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

$$\tilde{M}(\mathbf{f}) := M(\mathbf{f}) - M(\mathbf{0}), \tag{8}$$

which represents the flow-induced changes in the standard SDF.

Taking the difference between (4) and (5) and dividing element-wise by $\mathbf{P}(\mathbf{0})$, I obtain

$$\Delta\mathbf{p}(\mathbf{f}) = \mathbb{E}[\tilde{M}(\mathbf{f})\mathbf{R}_0]. \tag{9}$$

[13] 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)).

Since the fundamental return $\mathbf{R}_0$ is independent of flow $\mathbf{f}$, equation (9) implies that flow generates price impact only by varying the F-SDF $\tilde{M}(\mathbf{f})$. Intuitively, all the risk preferences of how the arbitrageur responds to flows are encoded by the F-SDF mapping,

$$\tilde{M}(\cdot): \quad \mathbf{f}(\omega_0) \in \mathbb{R}^N \longrightarrow \tilde{M}(\mathbf{f}(\omega_0)) \in L^2(\Omega_1, \mathcal{F}_1, \mathbb{P}_1), \tag{10}$$

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* $\tilde{M}(\mathbf{f})$ *satisfies*

$$\mathbb{E}[\tilde{M}(\mathbf{f})\mathbf{R}_0] = \Delta\mathbf{p}(\mathbf{f}), \tag{11}$$

$$\mathbb{E}[\tilde{M}(\mathbf{f})] = 0. \tag{12}$$

Equation (11) simply restates (9), which is a consequence of the LOOP. Equation (12) implies that the F-SDF has zero mean and follows from the assumption that the risk-free rate R_F does not depend on flow $\mathbf{f}$.

3.2 Portfolio Flow Theory

The standard portfolio theory focuses solely on the returns of the N assets. In contrast, my portfolio theory deals with two spaces—the return space and the flow space of the N assets. As a result, I need to establish an invariance principle to ensure that my portfolio flow theory is compatible with the existing portfolio return theory. To achieve this, I first derive the invariance principle and then apply it to construct portfolio flows.

Invariance principle. I contend that the inner product between flows and returns should remain invariant under portfolio formation. To understand why, consider that by absorbing flow $\mathbf{f}$ (in dollar amounts), the change in the arbitrageur's wealth is[14]

$$\Delta W(\mathbf{f}) = R_F \mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f}) - \mathbf{f}^\top (\mathbf{R}_0 - R_F \mathbf{1}). \tag{13}$$

The inner product between the flow and price impact, $\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f})$, represents the total compensation provided to the arbitrageur for absorbing the risk induced by the flow. The payoff risk is the inner product between the flow and excess fundamental return, $\mathbf{f}^\top (\mathbf{R}_0 - R_F \mathbf{1})$. Equation (13) also demonstrates why the flow $\mathbf{f}$ should be measured in dollar amounts, as the price impact $\Delta \mathbf{p}(\mathbf{f})$ and fundamental return $\mathbf{R}_0 - R_F \mathbf{1}$ are both measured per dollar.

Portfolio formation is essentially a change of basis used to study the same problem. Therefore, regardless of whether the problem is expressed in terms of the N original assets or N portfolios, the arbitrageur's wealth change $\Delta W(\mathbf{f})$ should remain the same. As per (13), in order to keep $\Delta W(\mathbf{f})$ invariant, it is necessary to maintain the invariance of both the payoff risk $\mathbf{f}^\top (\mathbf{R}_0 - R_F \mathbf{1})$ and total compensation $\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f})$. Since both $\mathbf{R}_0 - R_F \mathbf{1}$ and $\Delta \mathbf{p}(\mathbf{f})$ are returns, the principle is to preserve the inner product between flows and returns under portfolio formation.

Portfolio flows. I now apply the invariance principle to construct portfolio flows. First, I consider the straightforward case where one forms N portfolios using N assets. The $N \times N$ portfolio weight matrix is denoted as $\mathbf{B} = \{b_{n,k}\}$, where portfolio k holds $b_{n,k}$ dollars of asset n. The returns of the N assets are represented as $\mathbf{R} = (R_1, R_2, \ldots, R_N)^\top$. Standard portfolio theory implies that the returns of the N portfolios are $\mathbf{B}^\top \mathbf{R}$ (i.e., $\sum_{n=1}^{N} b_{n,k} R_n$). In

[14] To derive this equation, I express the flow in the unit of shares $\mathbf{h} = (f_1/P_1(\mathbf{0}), \ldots, f_N/P_N(\mathbf{0}))^\top$. The arbitrageur sells $\mathbf{h}$ units of shares at time 0 at price $\mathbf{P}(\mathbf{f})$. Consequently, $\Delta W(\mathbf{f}) = R_F \mathbf{h}^\top \mathbf{P}(\mathbf{f}) - \mathbf{h}^\top \mathbf{X}$, which simplifies to (13) using equations (6) and (7).

order to construct flows into the N portfolios, it is important to note that

$$\mathbf{f}^\top\mathbf{R} \neq (\mathbf{B}^\top\mathbf{f})^\top(\mathbf{B}^\top\mathbf{R}), \text{ but } \mathbf{f}^\top\mathbf{R} = (\underbrace{\mathbf{B}^{-1}\mathbf{f}}_{\text{portfolio flow}})^\top \underbrace{\mathbf{B}^\top\mathbf{R}}_{\text{portfolio return}}. \tag{14}$$

The first half of (14) asserts that constructing portfolio flows using $\mathbf{B}^\top\mathbf{f}$ (i.e., employing the same weighted-average method as for constructing portfolio returns) would alter the inner product between flows and returns, thus violating the invariance principle. The second half of (14) asserts that adhering to the invariance principle necessitates the construction of portfolio flows as

$$\mathbf{q} = \mathbf{B}^{-1}\mathbf{f} \text{ or, equivalently, } \mathbf{f} = \mathbf{B}\mathbf{q} = \sum_{k=1}^{N} \mathbf{b}_k q_k, \tag{15}$$

where $\mathbf{q} = (q_1, q_2, \ldots, q_N)^\top$ represents the flows into the N portfolios, and $\mathbf{b}_k = (b_{1,k}, b_{2,k}, \ldots, b_{N,k})^\top$ denotes the weights of portfolio k.

Equation (15) has an intuitive interpretation. I construct portfolio flows as *flows into portfolios*—portfolio flows q_k first enter portfolio k and are then allocated to individual assets based on portfolio weights $\mathbf{b}_k$. This clarifies why the N-asset flows conform to $\mathbf{f} = \sum_{k=1}^{N} \mathbf{b}_k q_k$. Consequently, to transition from asset flows to portfolio flows, one should apply the inverse of portfolio weights, $\mathbf{B}^{-1}$. On a deeper level, the invariance principle implies that the process of portfolio formation in the return space should be reversed in the flow space.

Equation (15) also uncovers a crucial connection—*portfolio weights* $\mathbf{B}$ *equal the betas of asset flows* $\mathbf{f}$ *to portfolio flows* $\mathbf{q}$. If portfolio flow q_k increases by \$1, the flow into asset n increases by \$$b_{n,k}$. As a result, the weight $b_{n,k}$ of asset n in portfolio k equals the beta of asset-n flow f_n with respect to portfolio-k flow q_k. This relationship plays a significant role when I later characterize cross-sectional price impacts and develop the factor model.

Second, I examine the general case in which $K \leq N$ portfolios are formed using N assets. I represent the $N \times K$ portfolio weight matrix as $\mathbf{B} = \{b_{n,k}\}$, where portfolio k holds $b_{n,k}$ dollars of asset n. Given that the number K of portfolios may be less than the number N of assets, it is unrealistic to expect a complete recovery of the inner product $\mathbf{f}^\top\mathbf{R}$ using the

K portfolios. Instead, the projection of the inner product onto the set of portfolios $\mathbf{B}$ could be kept invariant. To be specific, I obtain

$$\begin{aligned}(\overbrace{\mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{f}}^{\text{projected flow}})^\top \overbrace{\mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{R}}^{\text{projected return}} &= \mathbf{f}^\top\mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{R} \\ &= (\underbrace{(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{f}}_{\text{portfolio flow}})^\top \qquad \underbrace{\mathbf{B}^\top\mathbf{R}}_{\text{portfolio return}} . \end{aligned} \tag{16}$$

The flows $\mathbf{f}$ and returns $\mathbf{R}$ are first projected onto the portfolios $\mathbf{B}$, with $\mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top$ representing the standard projection matrix. The first equality stems from the fact that the projection matrix is idempotent. The second equality shows that the projected inner product can remain invariant under portfolio formations. When portfolio returns are defined in the standard manner, $\mathbf{B}^\top\mathbf{R}$, portfolio flows need to be defined as the *pseudoinverse*, $\mathbf{q} = (\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{f}$, which extends the special case $\mathbf{q} = \mathbf{B}^{-1}\mathbf{f}$ when $K = N$.

The pseudoinverse can also be interpreted as a cross-sectional regression of asset flows $\mathbf{f}$ on portfolio weights $\mathbf{B}$. Observe that if portfolio-k flow increases by \$1, asset-n flow rises by \$$b_{n,k}$. Consequently, if the K portfolio flows are $q_1, q_2, \ldots, q_K$, the resulting flow into asset n is $\sum_{k=1}^K b_{n,k}q_k$. To find the portfolio flows $q_1, q_2, \ldots, q_K$ that best approximate the observed asset flows $\mathbf{f} = (f_1, f_2, \ldots, f_N)^\top$, I project asset flows f_n onto portfolio weights $b_{n,1}, b_{n,2}, \ldots, b_{n,K}$, which yields the pseudoinverse[15] $\mathbf{q} = (\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{f}$.

One may think that the pseudoinverse $(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{f}$ and the transpose $\mathbf{B}^\top\mathbf{f}$ would be identical if portfolio weights are orthonormal (i.e., $\mathbf{B}^\top\mathbf{B} = \mathbf{I}$), implying that the discrepancy between the correct and incorrect constructions is minimal. However, orthonormal portfolio weights are economically implausible. For instance, the correlation between the portfolio weights of the Fama and French (1993) market and small-minus-big factors is approximately -0.8. This is because the Fama-French construction aims to orthogonalize portfolio returns

[15] The cross-sectional regression could alternatively be run as weighted least squares (WLS) to obtain $\mathbf{q} = (\mathbf{B}^\top\mathbf{W}\mathbf{B})^{-1}\mathbf{B}^\top\mathbf{W}\mathbf{f}$ for some $N \times N$ weighting matrix $\mathbf{W}$. Online Appendix B considers this case. I demonstrate that the WLS construction can be interpreted as first rotating the N asset flows $\mathbf{f}$ into N equivalent portfolios and then projecting these N portfolios onto the K portfolio weights $\mathbf{B}$.

(which possess economic significance) rather than orthogonalize portfolio weights (which lack such meaning).

3.3 Flow-Based Hansen-Jagannathan Bound

I establish the flow-based Hansen-Jagannathan bound, which states that for any fixed flow, the minimum volatility of the F-SDF is equal to the maximum price impact ratio. This bound encapsulates the notion that "noisy flows shock the SDF and change asset prices today, leading to flow-induced return predictability in the future."

By the pricing equation (11), the price impact of any portfolio $\mathbf{c} \in \mathbb{R}^N$ is $\mathbf{c}^\top \Delta \mathbf{p}(\mathbf{f}) = \mathbb{E}[\mathbf{c}^\top \mathbf{R}_0 \tilde{M}(\mathbf{f})]$. Using the F-SDF's zero-mean property from (12), I have

$$\mathbb{E}[\mathbf{c}^\top \mathbf{R}_0 \tilde{M}(\mathbf{f})] = \operatorname{corr}(\mathbf{c}^\top \mathbf{R}_0, \tilde{M}(\mathbf{f}))\sigma(\mathbf{c}^\top \mathbf{R}_0)\sigma(\tilde{M}(\mathbf{f})), \tag{17}$$

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

$$\sigma(\tilde{M}(\mathbf{f})) \geq \max_{\mathbf{c} \in \mathbb{R}^N} \frac{\mathbf{c}^\top \Delta \mathbf{p}(\mathbf{f})}{\sigma(\mathbf{c}^\top \mathbf{R}_0)}. \tag{18}$$

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 (18) is the *maximum price impact ratio (MPIR)* across all portfolios.

Inequality (18) demonstrates that, for any fixed flow $\mathbf{f}$, the F-SDF's volatility is no less than the maximum price impact ratio. Appendix A.1 establishes that this bound is tight.

PROPOSITION 2. *The flow-based Hansen-Jagannathan bound is*

$$\min_{\tilde{M}(\mathbf{f})} \sigma(\tilde{M}(\mathbf{f})) = \max_{\mathbf{c} \in \mathbb{R}^N} \frac{\mathbf{c}^\top \Delta \mathbf{p}(\mathbf{f})}{\sigma(\mathbf{c}^\top \mathbf{R}_0)} \tag{19}$$

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

$$\tilde{M}^*(\mathbf{f}) = \mathbf{\Delta p}(\mathbf{f})^\top \mathrm{var}(\mathbf{R}_0)^{-1}(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]). \tag{20}$$

The crux of Proposition 2 lies in the portfolio that generates the maximum price impact for a given level of fundamental risk. Intuitively, shorting the MPIR portfolio represents the mean-variance optimal strategy to capitalize on flow-induced changes in risk premiums. Equation (19) shows that the MPIR is equal to the minimum volatility of the F-SDF, which corresponds to the flow-induced changes in standard SDFs.

The minimum volatility F-SDF $\tilde{M}^*(\mathbf{f})$ in (20) represents the projection of any F-SDF onto the space spanned by $\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]$. I denote this space as $\underline{\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]}$ and refer to it as the fundamental-risk space. Appendix A.1 provides a proof for the following corollary.

COROLLARY 1. *For any flow* $\mathbf{f}$, *the minimum-volatility F-SDF* $\tilde{M}^*(\mathbf{f})$ *in* (20) *is the unique F-SDF in the fundamental-risk space* $\underline{\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]}$.

Given this corollary, I require that $\tilde{M}(\mathbf{f}) \in \underline{\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]}$ for every $\mathbf{f} \in \mathbb{R}^N$ in order to streamline the characterization of the F-SDF mapping from $\mathbf{f}$ to $\tilde{M}(\mathbf{f})$.

Relationship with standard theory. The original Hansen and Jagannathan (1991) bound demonstrates that the maximum Sharpe ratio equals the minimum volatility of the standard SDF (multiplied by the risk-free rate R_F). Applied to my setting and utilizing (6) and (7), the minimum-volatility standard SDF under flow $\mathbf{f}$ is[16]

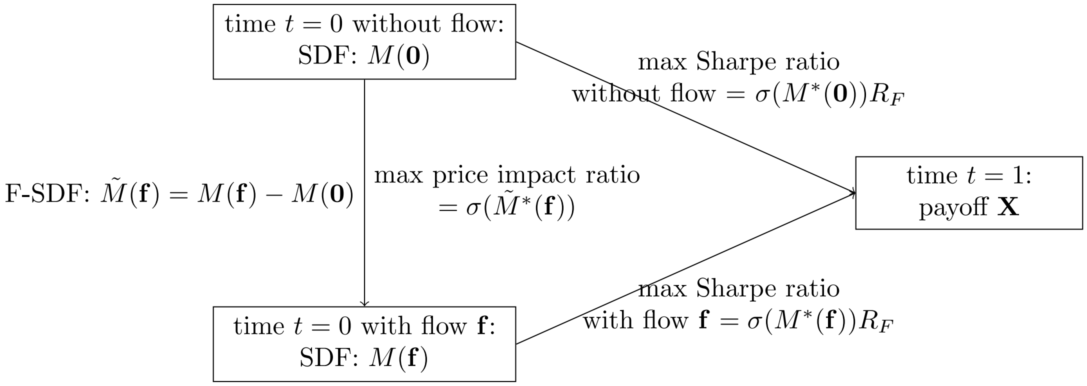
$$M^*(\mathbf{f}) = \frac{1}{R_F} + \left(\mathbf{\Delta p}(\mathbf{f}) - \frac{\mathbb{E}[\mathbf{R}_0 - R_F\mathbf{1}]}{R_F}\right)^\top \mathrm{var}(\mathbf{R}_0)^{-1}(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]). \tag{21}$$

Therefore, the difference between two minimum-volatility standard SDFs, $M^*(\mathbf{f})$ and $M^*(\mathbf{0})$, is equal to my minimum-volatility F-SDF, $\tilde{M}^*(\mathbf{f})$.

Figure 2 illustrates the relationship with standard theory. The original Hansen-Jagannathan bound demonstrates that the volatility of $M^*(\mathbf{f})$ and $M^*(\mathbf{0})$ (times the risk-free rate R_F)

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

Figure 2. Relationship with standard theory


time $t = 0$ without flow:
SDF: $M(\mathbf{0})$
max Sharpe ratio
without flow $= \sigma(M^*(\mathbf{0}))R_F$
F-SDF: $\tilde{M}(\mathbf{f}) = M(\mathbf{f}) - M(\mathbf{0})$
max price impact ratio
$= \sigma(\tilde{M}^*(\mathbf{f}))$
time $t = 1$:
payoff $\mathbf{X}$
max Sharpe ratio
with flow $\mathbf{f} = \sigma(M^*(\mathbf{f}))R_F$
time $t = 0$ with flow $\mathbf{f}$:
SDF: $M(\mathbf{f})$

Notes: The F-SDF, $\tilde{M}(\mathbf{f})$, is the difference between two standard SDFs, $M(\mathbf{f})$ and $M(\mathbf{0})$. The original Hansen-Jagannathan bound demonstrates 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 without flow $\mathbf{f}$, respectively. The flow-based Hansen-Jagannathan bound indicates that the volatility of $\tilde{M}^*(\mathbf{f}) = M^*(\mathbf{f}) - M^*(\mathbf{0})$ corresponds to the maximum price impact ratio under flow $\mathbf{f}$.

corresponds to the respective maximum Sharpe ratios. The flow-based Hansen-Jagannathan bound shows that the volatility of $\tilde{M}^*(\mathbf{f}) = M^*(\mathbf{f}) - M^*(\mathbf{0})$ equals the MPIR.

Moreover, An, Su, and Wang (2022) show that a simple transformation exists between the optimal portfolios depicted in Figure 2. Specifically, the portfolio achieving the maximum Sharpe ratio in the economy with flow $\mathbf{f}$ is obtained by longing the portfolio that achieves the maximum Sharpe ratio in the economy without flow and shorting the MPIR portfolio. In other words, to achieve the maximum Sharpe ratio, an investor simply needs to first estimate the MPIR portfolio (which encapsulates all the flow information) and then combine it with fundamental investing (which does not require any flow information). Empirically, they estimate the MPIR portfolio using mutual fund flows. They find that the MPIR strategy enhances the out-of-sample Sharpe ratio of a broad spectrum of firm characteristics-based anomaly portfolios (which proxy fundamental investing) by an average of 0.3 per annum.

3.4 Linear Model of F-SDF

The results so far have assumed only the LOOP. I now introduce additional economic constraints to further refine the form of the F-SDF.

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 have

$$\Delta\mathbf{p}(a_1\mathbf{f}_1 + a_2\mathbf{f}_2) = a_1\Delta\mathbf{p}(\mathbf{f}_1) + a_2\Delta\mathbf{p}(\mathbf{f}_2). \tag{22}$$

The linearity of price impact is an assumption, not a consequence of the LOOP. The economy with flow $\mathbf{f}_1$ differs from the economy with flow $\mathbf{f}_2$. The LOOP does not apply between the two economies. Linearity implies that price impact doubles if flow doubles and that price impact is additive for two flows. The linearity assumption is made for two reasons. Theoretically, quadratic-normal models in the literature fall within this broader class of linear price impact models. Empirically, researchers implicitly linearize price impacts, at least locally, when estimating price elasticities.

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

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

Recall that (13) demonstrates that $\mathbf{f}^\top\Delta\mathbf{p}(\mathbf{f})$ represents the arbitrageur's risk compensation for absorbing flow $\mathbf{f}$. Assumption 2 stipulates that this compensation must be strictly positive for any non-zero flow. Intuitively, $\mathbf{f}^\top\Delta\mathbf{p}(\mathbf{f}) > 0$ indicates that when there is an inflow into an asset, the asset's price should increase rather than decrease.

The LOOP, linearity, and positive compensation for risk enable me to express the F-SDF in a linear form. Appendix A.2 contains a proof.

PROPOSITION 3. *There exists a unique F-SDF that satisfies the following:*

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

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

iii. projection: $\tilde{M}(\mathbf{f}) \in \underline{\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]}$ *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

$$\tilde{M}^*(\mathbf{f}) = (\mathbf{Y}\mathbf{f})^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]), \tag{23}$$

for some $N \times N$ matrix $\mathbf{Y}$ such that $\mathrm{var}(\mathbf{R}_0)\mathbf{Y}$ is positive definite.

The F-SDF $\tilde{M}^*(\mathbf{f})$ in (23) has N^2 free parameters $\mathbf{Y}$. To simplify the notation, in the following discussion, I use $\tilde{M}(\mathbf{f})$ instead of $\tilde{M}^*(\mathbf{f})$ to represent this unique F-SDF.

Roadmap for the remaining paper. The F-SDF in (23) implies a price impact model

$$\Delta\mathbf{p}(\mathbf{f}) = \mathbb{E}[\tilde{M}(\mathbf{f})\mathbf{R}_0] = \mathrm{var}(\mathbf{R}_0)\mathbf{Y}\mathbf{f}. \tag{24}$$

Intuitively, equation (24) contains the right components, as $\mathrm{var}(\mathbf{R}_0)$ represents risk and $\mathbf{f}$ corresponds to flow. However, the N^2 free parameters in $\mathbf{Y}$ and the requirement for $\mathrm{var}(\mathbf{R}_0)\mathbf{Y}$ to be positive definite impede both theoretical interpretation and empirical application. Consequently, the remainder of this paper introduces additional economic constraints to reduce the number of free parameters.

Specifically, under model (24), the flow into any asset can impact the price of any other asset or portfolio. Intuition suggests that some price impacts in the cross-section are plausible, while others are not. This cross-sectional restriction on price impacts is formalized as the Irrelevance of Uncorrelated Flows (IUF), which reduces the number of free parameters in (24) from N^2 to N. I then impose the flow-based APT to further reduce this number from N to a low-dimensional K. The IUF is an economically novel and somewhat abstract concept that employs an unfamiliar mathematical tool. Thus, the remainder of this section and Section 3.6 aim to help readers understand various aspects of the IUF. Section 3.7 presents the flow-based APT.

Another way to comprehend the IUF is by comparing it with the literature's quadratic-normal setup. Section 3.5 demonstrates that the quadratic-normal setup, which directly reduces the free parameters of (24) from N^2 to 1, is a special case of the IUF. Notably, my more general model also unveils a new channel of cross-impacts.

Irrelevance of Uncorrelated Flows. Intuitively, not all cross-asset price impacts are plausible. For a simple example, consider two assets that have uncorrelated flows $\mathbf{f}$ and uncorrelated fundamental returns $\mathbf{R}_0$. In this case, one should not expect the flow into the first asset to impact the second asset's price. The IUF formalizes this intuition by eliminating all such implausible price impacts in a setting with general covariance structures of flows and fundamental returns.

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

I define the set of matrices,

$$\mathcal{C} := \left\{ \mathbf{C} \in \mathbb{R}^{N \times N} \middle| \, \mathrm{var}(\mathbf{R}_0)\mathrm{var}(\mathbf{f})\mathbf{C} = \mathbf{C}\mathrm{var}(\mathbf{f})\mathrm{var}(\mathbf{R}_0) \right\}. \tag{25}$$

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

$$\mathbf{d} = \left(\mathrm{cov}(f_1, s), \mathrm{cov}(f_2, s), \ldots, \mathrm{cov}(f_N, s)\right)^\top. \tag{26}$$

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

From a big-picture perspective, the IUF asserts that some types of price impacts in the cross-section should be zero. Matrix math is employed to precisely characterize these price impacts. Let me explain the details.

In the cross-section of the N assets, the key objects are the covariance structure of fundamental returns $\mathrm{var}(\mathbf{R}_0)$ and noisy flows $\mathrm{var}(\mathbf{f})$. Collectively, they determine the flow-induced fundamental risk that the arbitrageur absorbs. Intuitively, any $N \times N$ matrix $\mathbf{C}$ that satisfies equation (25) selects one particular combination of the flow-induced risk. The set $\mathcal{C}$ characterizes all possible combinations of the flow-induced risk. Mathematically, this occurs because commuting matrices preserve eigenvectors, which represent the most significant directions of risks induced by the flows.

The IUF's objective is to characterize the relationship between the price impact of any portfolio $\mathbf{a} \in \mathbb{R}^N$ and flow s. To achieve this, I need to identify the portfolio that s flows into. In other words, if flow s varies, how would the flow into each asset n change? The answer is the beta of asset flow f_n to flow s, given by $\text{cov}(f_n, s)/\text{var}(s)$. Recall the key insight from my portfolio theory (15) that flow betas are portfolio weights. Therefore, the weight of asset n in the portfolio that s flows into is $\text{cov}(f_n, s)/\text{var}(s)$. This result gives the portfolio $\mathbf{d}$ in (26), where, for maximal simplicity, I further normalize all portfolio weights by $\text{var}(s)$.

In plain words, the condition "$\mathbf{a}^\top \mathbf{C} \mathbf{d} = 0$ for all $\mathbf{C} \in \mathcal{C}$" implies that the portfolios $\mathbf{a}$ and $\mathbf{d}$ are uncorrelated for every combination of the flow-induced risk. Under this condition, the IUF mandates that flow s should not impact the price of portfolio $\mathbf{a}$, which is characterized through covariance $\text{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\mathbf{f}), s) = 0$.

The following Theorem 1 presents the linear model of the F-SDF.

THEOREM 1. *There exists a unique F-SDF $\tilde{M}(\cdot)$ that satisfies all restrictions in Proposition 3 and the IUF Assumption 3.*

The proof of Theorem 1 is provided in Appendix A.4. I now present the key steps. Using Proposition 3, I explicitly construct a particular F-SDF of form (23), referred to as the canonical form. I then prove that the price impacts $\Delta \mathbf{p}(\mathbf{f})$ satisfy the IUF if and only if the corresponding F-SDF $\tilde{M}(\mathbf{f})$ can be expressed in the canonical form.

Canonical form of F-SDF. I introduce a technical restriction to simplify the exposition in the main text.

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* $\text{var}(\mathbf{U}\mathbf{f})$ *has distinct eigenvalues.*

Intuitively, $\mathbf{U}\mathbf{f}$ represents the fundamental-risk-weighted flow, as $\mathbf{U}$ is derived from the

Cholesky decomposition of the fundamental risk matrix $\mathrm{var}(\mathbf{R}_0)$. As I shall demonstrate, $\mathbf{Uf}$ plays a critical role in determining the arbitrageur's price response to absorbing the flow $\mathbf{f}$. Recall that in undergraduate linear algebra courses, distinct eigenvalues are the straightforward and general case, while duplicate eigenvalues are the complex and knife-edge case. The same applies to my theory. Online Appendix A presents the theory when $\mathrm{var}(\mathbf{Uf})$ can have duplicate eigenvalues. In that scenario, the IUF does not constrain how flows impact the price of portfolios within the same eigenspace. Intuitively, 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 constructs the orthogonalized factors for the canonical form. Appendix A.3 provides a proof.

LEMMA 1. *There exist N unique (up to multiplication by -1) portfolios $\mathbf{b}_n = (b_{1,n}, b_{2,n}, \ldots, b_{N,n})^\top$, which I denote as a matrix $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N)$, satisfying*

i. factor decomposition of flow

$$\mathbf{f} = \sum_{n=1}^{N} \mathbf{b}_n q_n = \mathbf{B}\mathbf{q}, \tag{27}$$

where q_n is the flow into portfolio n and $\mathbf{q} = (q_1, q_2, \ldots, q_N)^\top$.

ii. uncorrelated fundamental risk

$$\mathbf{B}^\top \mathrm{var}(\mathbf{R}_0)\mathbf{B} = \mathbf{I}_N. \tag{28}$$

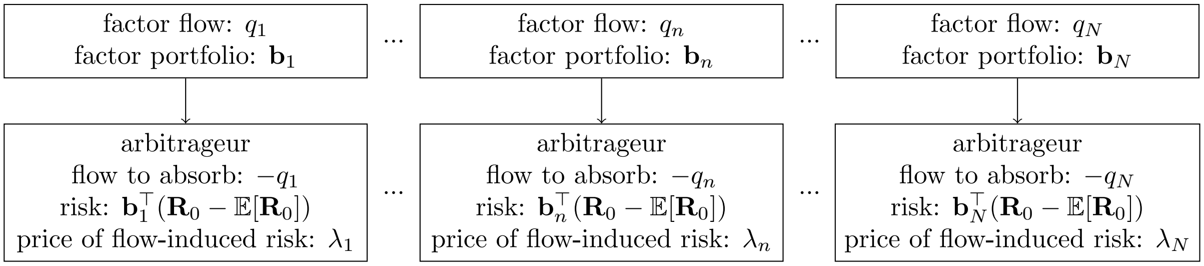
iii. uncorrelated flow

$$\mathrm{var}(\mathbf{q}) = \mathrm{diag}(\pi_1, \pi_2, \ldots, \pi_N) := \mathbf{\Pi}, \tag{29}$$

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 *(orthogonalized) factor portfolios* and the flows

Figure 3. Intuition for the canonical form of the F-SDF


factor flow: q_1
factor portfolio: $\mathbf{b}_1$
...
factor flow: q_n
factor portfolio: $\mathbf{b}_n$
...
factor flow: q_N
factor portfolio: $\mathbf{b}_N$
arbitrageur
flow to absorb: $-q_1$
risk: $\mathbf{b}_1^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])$
price of flow-induced risk: λ_1
...
arbitrageur
flow to absorb: $-q_n$
risk: $\mathbf{b}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])$
price of flow-induced risk: λ_n
...
arbitrageur
flow to absorb: $-q_N$
risk: $\mathbf{b}_N^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])$
price of flow-induced risk: λ_N

Notes: The N orthogonalized factor portfolios exhibit uncorrelated fundamental risk, $\mathbf{b}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])$, and uncorrelated flow, q_n. Factor portfolios do not have cross-impacts, and the arbitrageur responds to each factor flow independently. Each factor's price of flow-induced risk is a free parameter, denoted as $\lambda_n > 0$.

$q_1, q_2, \ldots, q_N$ as *(orthogonalized) factor flows*. The factor decomposition (27) justifies this interpretation. According to (28), factor portfolios have uncorrelated and unit fundamental risk. As per (29), factor flows are uncorrelated and ranked from the largest variance to the smallest. These factor portfolios are uniquely determined.

To better comprehend the orthogonalized factors, I compare them to a naive principal component analysis (PCA) performed on the flow $\mathbf{f} = \mathbf{B}\mathbf{q}$. The naive PCA enforces the orthogonalization condition $\mathbf{B}^\top\mathbf{B} = \mathbf{I}$, which implies that the portfolio weights of different factors are orthogonal to one another. Regrettably, orthogonal portfolio weights lack economic significance. Instead, my orthogonalized factors are constructed under the condition $\mathbf{B}^\top\text{var}(\mathbf{R}_0)\mathbf{B} = \mathbf{I}$, signifying that the fundamental risks of different factors are orthogonal to each other. The corresponding factor flows $q_1, q_2, \ldots, q_N$ are the principal components of the fundamental-risk-weighted flows $\mathbf{U}\mathbf{f}$, where $\text{var}(\mathbf{R}_0) = \mathbf{U}^\top\mathbf{U}$, and $\pi_1, \pi_2, \ldots, \pi_N$ are the eigenvalues of $\text{var}(\mathbf{U}\mathbf{f})$. The orthogonalized factors yield the canonical form of the F-SDF.

DEFINITION 1. *The canonical form of the F-SDF is*

$$\tilde{M}(\mathbf{f}) = \sum_{n=1}^{N} \lambda_n q_n \mathbf{b}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) = (\boldsymbol{\Lambda}\mathbf{q})^\top\mathbf{B}^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]), \tag{30}$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ *is an* $N \times N$ *diagonal matrix with* $\lambda_n > 0$ *for all* n.

Figure 3 illustrates the intuitions underlying the canonical form. For each factor n,

$\mathbf{b}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])$ represents the per-unit fundamental risk when the arbitrageur absorbs the flow q_n. The N orthogonalized factor portfolios have uncorrelated fundamental risk, $\mathbf{b}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])$, and uncorrelated flow, q_n. The IUF implies that these factor portfolios do not have cross-impacts, and the arbitrageur reacts to each factor flow independently. For each factor n, the price change for each unit of factor risk, $\mathbf{b}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])$, induced by factor flow q_n, is measured by a free parameter, λ_n, which I term *the price of flow-induced risk.* The intuition that flows cause positive price impacts is characterized by the restriction $\lambda_n > 0$.

Comparing the canonical form (30) with the linear form (23), one can see that the IUF restricts the price-of-flow-induced-risk matrix, $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_N)$, to be diagonal. This means that cross-impacts do not exist between any two orthogonalized factors $n \neq m$, and thus the canonical form does not have any off-diagonal terms of the form $q_n \mathbf{b}_m^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])$.

Appendix A.4 formally proves that price impacts $\Delta\mathbf{p}(\mathbf{f})$ satisfy the IUF *if and only if* the corresponding F-SDF $\tilde{M}(\mathbf{f})$ can be written into the canonical form (30), thereby completing the proof of Theorem 1. In other words, the canonical form (30) completely characterizes the IUF condition and vice versa.

Structural restrictions for the general form of the F-SDF. Empirically, one might not want to use the orthogonalized factor portfolios $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_N)$ to characterize the F-SDF and the cross-section of price impacts. Instead, one might want to choose a set of portfolios $\tilde{\mathbf{B}} = (\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \dots, \tilde{\mathbf{b}}_N)$ 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 $\tilde{\mathbf{q}} = (\tilde{q}_1, \tilde{q}_2, \dots, \tilde{q}_N)^\top$ 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.

In the general case, the price-of-flow-induced-risk matrix is not necessarily diagonal but satisfies a structural restriction imposed by the IUF. Such a matrix also has N degrees of freedom, as the canonical form does. Appendix A.5 provides a proof.

PROPOSITION 4. *The general form of the F-SDF is*

$$\tilde{M}(\mathbf{f}) = \sum_{n=1}^{N}\sum_{m=1}^{N} \tilde{\lambda}_{n,m}\tilde{q}_m \tilde{\mathbf{b}}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) = (\tilde{\mathbf{\Lambda}}\tilde{\mathbf{q}})^\top \tilde{\mathbf{B}}^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]). \tag{31}$$

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

$$\mathbf{O}^{-1}\tilde{\mathbf{\Lambda}}\mathbf{O} = \mathbf{\Lambda}, \textit{ where } \mathbf{\Lambda} \textit{ is diagonal and positive definite,} \tag{32}$$

$$\mathbf{O}^\top\tilde{\mathbf{B}}^\top \mathrm{var}(\mathbf{R}_0)\tilde{\mathbf{B}}\mathbf{O} = \mathbf{I}_N, \tag{33}$$

$$\mathbf{O}\mathbf{\Pi}\mathbf{O}^\top = \mathrm{var}(\tilde{\mathbf{q}}), \textit{ where } \mathbf{\Pi} \textit{ is diagonal and positive definite.} \tag{34}$$

Linear model of price impacts. By the canonical form (30), the cross-section of price impacts satisfies

$$\Delta\mathbf{p}(\mathbf{f}) = \mathbb{E}[\tilde{M}(\mathbf{f})\mathbf{R}_0] = \sum_{n=1}^{N} \lambda_n q_n \mathrm{cov}(\mathbf{R}_0, \mathbf{b}_n^\top \mathbf{R}_0). \tag{35}$$

In particular, the price impact of the factor portfolio $\mathbf{b}_n$ is

$$\mathbf{b}_n^\top \Delta\mathbf{p}(\mathbf{f}) = \lambda_n q_n \mathrm{var}(\mathbf{b}_n^\top \mathbf{R}_0). \tag{36}$$

My model is similar to the canonical factor model. The quantity of risk induced by every dollar of factor-n flow q_n is the factor's fundamental risk $\mathrm{var}(\mathbf{b}_n^\top \mathbf{R}_0)$. The price of flow-induced risk λ_n measures how each unit of factor risk induced by factor flow changes the factor price. In the next section, I provide more intuitions on λ_n by comparing it with equilibrium models.

Similarly, the general form of the F-SDF in (31) implies

$$\Delta\mathbf{p}(\mathbf{f}) = \sum_{n=1}^{N}\sum_{m=1}^{N} \tilde{\lambda}_{n,m}\tilde{q}_m \mathrm{cov}(\mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0), \tag{37}$$

where $\tilde{\lambda}_{n,m}$ measures how flow $\tilde{q}_m$ impacts the price of portfolio $\tilde{\mathbf{b}}_n$. Recall that the set of parameters $\tilde{\lambda}_{n,m}$ are not N^2 free parameters but have only N degrees of freedom as determined by Proposition 4.

3.5 Commonality-In-Flow Paradox

The standard approach to solving multi-asset price impact models involves imposing assumptions on both the utility function and the payoff distribution. Commonly, the utility function is assumed to be quadratic or CARA, while the payoff distribution is assumed to be multivariate normal. Online Appendix C shows that the static CARA (quadratic)-normal price impact model is a special case of my IUF model (35). This is done by further requiring that $\lambda_1 = \lambda_2 = \cdots = \lambda_N = \gamma/(\mu R_F)$, where γ represents the arbitrageurs' CARA risk aversion, μ denotes the mass of arbitrageurs, and R_F is the gross risk-free rate.

Economically, the static quadratic-normal setup requires that arbitrageurs must be equally averse to risks induced by flows into different portfolios. This restriction is strong and unrealistic, which is why my IUF condition removes it. Simultaneously, the static quadratic-normal setup also imposes a realistic restriction on the cross-section of price impacts, which is precisely revealed by my IUF condition.

In addition to providing greater flexibility than the quadratic-normal model, my model also uncovers a new channel of cross-asset price impacts. I illustrate this channel using a commonality-in-flow paradox—do portfolios with uncorrelated fundamental returns have any cross-impact other than the mechanical effect driven by the flow covariance? My model predicts in general yes, and the static quadratic-normal model predicts absolutely no.

The new channel arises precisely because my model has more degrees of freedom in λ_n than the static quadratic-normal model. In my model, the flow of asset B changes the prices of different factors, which then affect the price of asset A by no arbitrage. In the special quadratic-normal setup where all factors have the same λ_n, the changing prices of different factors exactly cancel out, leaving the price of asset A unchanged.

I now present the precise math. I consider N portfolios $\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N$ with uncorrelated fundamental returns, i.e., $\mathrm{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ for $n \neq m$. 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 $\tilde{q}_n$.

By (37), the price impact of portfolio $\tilde{\mathbf{b}}_n$ under my IUF model is

$$\tilde{\mathbf{b}}_n^\top \Delta\mathbf{p}(\mathbf{f}) = \sum_{m=1}^{N} \tilde{\lambda}_{n,m} \tilde{q}_m \mathrm{var}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0), \tag{38}$$

where the price-of-flow-induced-risk matrix $\tilde{\mathbf{\Lambda}} = \{\tilde{\lambda}_{n,m}\}$ is given by equation (32),

$$\tilde{\mathbf{\Lambda}} = \mathbf{O}\mathbf{\Lambda}\mathbf{O}^{-1} = \mathbf{O}\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)\mathbf{O}^{-1}. \tag{39}$$

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. Equation (39) shows that the off-diagonal term $\tilde{\lambda}_{n,m}$ is generally nonzero for $m \neq n$. Therefore, by (38), 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, which further requires $\lambda_1 = \lambda_2 = \cdots = \lambda_N$, is a special case of my IUF model. In this special case, equation (39) implies that

$$\tilde{\mathbf{\Lambda}} = \mathbf{O}\mathrm{diag}(\lambda_1, \lambda_1, \ldots, \lambda_1)\mathbf{O}^{-1} = \mathrm{diag}(\lambda_1, \lambda_1, \ldots, \lambda_1) \tag{40}$$

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.

Equation (39) also implies a powerful empirical test for the new channel of cross-impacts. If the prices of flow-induced risk $\lambda_1, \lambda_2, \ldots, \lambda_N$ are not equal, one can construct portfolios $\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N$ that have uncorrelated fundamental returns but cross-impacts between each other. Empirically, An, Su, and Wang (2022) reject the null hypothesis that all λ_n are equal and thus support the new channel.

3.6 Optimality of the IUF Orthogonalization

This section proves that the IUF is the optimal orthogonalization—the IUF is the only way of reducing the N^2 model (24) to an N-factor model, while keeping each factor's price elasticity invariant and keeping the arbitrageur's expected utility invariant. Simply put, if one wants to choose N factors to span the cross-section of price impacts as in (35), the N factors selected by my IUF condition are the unique optimal choice. This section is presented under the regularity Assumption 4. Online Appendix A presents the general theory without the regularity assumption.

I first define the meaning of orthogonalizing the N^2 model to an N-factor model.

DEFINITION 2. *Each set of 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 (24) under portfolios $\tilde{\mathbf{B}}$ is*[17]

$$\Delta\mathbf{p}(\mathbf{f}) = \sum_{n=1}^{N}\sum_{m=1}^{N} \tilde{\lambda}_{n,m}\tilde{q}_m \mathrm{cov}(\mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0), \tag{41}$$

with the $N \times N$ free parameters $\tilde{\mathbf{\Lambda}} = \{\tilde{\lambda}_{n,m}\}$ and portfolio flows $\tilde{\mathbf{q}} = (\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_N)^\top$. The orthogonalized N-factor model under portfolios $\tilde{\mathbf{B}}$ is defined as

$$\Delta\bar{\mathbf{p}}(\mathbf{f}) = \sum_{n=1}^{N} \tilde{\lambda}_{n,n}\tilde{q}_n \mathrm{cov}(\mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0). \tag{42}$$

By Definition 2, model orthogonalization means that one picks a specific set of N portfolios $\tilde{\mathbf{B}} = (\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N)$ and removes the off-diagonal terms of the corresponding $\tilde{\mathbf{\Lambda}} = \{\tilde{\lambda}_{n,m}\}$. By Theorem 1, the IUF is a particular model orthogonalization that chooses the specific factor portfolios $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N)$ with uncorrelated fundamental returns and uncorrelated flows.

To understand why the IUF is the optimal orthogonalization, I next define the arbitrageur's expected utility. Let the arbitrageur's wealth under flow $\mathbf{f}$ be denoted as $W(\mathbf{f})$.

[17] The transformation between (24) and (41) is given by $\mathbf{Y} = \tilde{\mathbf{B}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{B}}^{-1}$.

Also, let the arbitrageur's utility function on her wealth be denoted as $u(\cdot)$. By absorbing flow $\mathbf{f}$, the arbitrageur's expected utility is given by

$$\mathbb{E}[u(W(\mathbf{f}))] = \mathbb{E}[u(W(\mathbf{0}) + R_F \mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f}) - \mathbf{f}^\top(\mathbf{R}_0 - R_F \mathbf{1}))], \tag{43}$$

where I use equation (13). As discussed in Section 3.2, $\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})$ represents the arbitrageur's compensation for absorbing the flow-induced risk, and $\mathbf{f}^\top(\mathbf{R}_0 - R_F \mathbf{1})$ represents the total payoff risk.

Since flow $\mathbf{f}$ is random, the compensation for risk, $\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})$, also varies. However, flow $\mathbf{f}$ is independent of the assets' payoff $\mathbf{X}$, implying that $\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})$ is also independent of $\mathbf{X}$. Therefore, 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 risk, i.e., for some function $\tilde{u}(\cdot)$,

$$u(W(\mathbf{0}) + R_F \mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f}) - \mathbf{f}^\top(\mathbf{R}_0 - R_F \mathbf{1})) = R_F \mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f}) + \tilde{u}(W(\mathbf{0}) - \mathbf{f}^\top(\mathbf{R}_0 - R_F \mathbf{1})). \tag{44}$$

Using equations (43) and (44), I simplify the arbitrageur's expected utility as

$$\mathbb{E}[u(W(\mathbf{f}))] = R_F \mathbb{E}[\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})] + \mathbb{E}[\tilde{u}(W(\mathbf{0}) - \mathbf{f}^\top(\mathbf{R}_0 - R_F \mathbf{1}))]. \tag{45}$$

The key insight from (45) 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 for risk $\mathbb{E}[\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})]$. Therefore, if I hold $\mathbb{E}[\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})]$ 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.6 provides a proof.

THEOREM 2. *Fix any model orthogonalization* $\tilde{\mathbf{B}} = (\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N)$. *The orthogonalized model (42) 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* $\tilde{\boldsymbol{\Lambda}} = \{\tilde{\lambda}_{n,m}\}$,

- *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 (41) and the orthogonalized model (42),*

$$\frac{\partial \tilde{\mathbf{b}}_n^\top \Delta \mathbf{p}(\mathbf{f})}{\partial \tilde{q}_n} = \frac{\partial \tilde{\mathbf{b}}_n^\top \Delta \bar{\mathbf{p}}(\mathbf{f})}{\partial \tilde{q}_n}. \tag{46}$$

- *the arbitrageur's expected utility $\mathbb{E}[u(W(\mathbf{f}))]$ remains invariant under (41) and (42).*

Theorem 2 is an *if-and-only-if* characterization, demonstrating that the IUF is the only model orthogonalization that satisfies two key properties. First, (46) ensures that each factor's price elasticity 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 elasticity $\tilde{\lambda}_{n,n}$.

Second, model orthogonalization eliminates off-diagonal terms of $\tilde{\mathbf{\Lambda}} = \{\tilde{\lambda}_{n,m}\}$, 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 $\mathbb{E}[u(W(\mathbf{f}))]$. My arbitrage approach to price impacts does not make any parametric assumptions regarding the arbitrageur's utility function. Nonetheless, I prove that the IUF ensures that the expected utility remains invariant under the N^2 model and the orthogonalized N-factor model.

3.7 Flow-Based APT

This section develops the flow-based APT to reduce the N-factor model of price impacts (35) to a low-dimensional K-factor model.

I write the N-asset flow as the N-orthogonalized-factor structure $\mathbf{f} = \sum_{n=1}^{N} \mathbf{b}_n q_n$. The factor portfolios $\mathbf{b}_n$ and factor flows q_n are the orthogonalized ones in the canonical form (30), and flows q_n are sorted such that their variance satisfies $\pi_1 > \pi_2 > \cdots > \pi_N > 0$. The N-factor price impact model (35) is $\Delta \mathbf{p}(\mathbf{f}) = \sum_{n=1}^{N} \lambda_n q_n \mathrm{cov}(\mathbf{R}_0, \mathbf{b}_n^\top \mathbf{R}_0)$. In the flow-based APT, I use the first K orthogonalized factor flows, $\mathbf{f} = \sum_{k=1}^{K} \mathbf{b}_k q_k + \mathbf{e}$, where $\mathbf{e}$ is an $N \times 1$

vector of idiosyncratic flows. The corresponding K-factor price impact model is

$$\Delta\check{\mathbf{p}}(\mathbf{f}) = \sum_{k=1}^{K} \lambda_k q_k \mathrm{cov}(\mathbf{R}_0, \mathbf{b}_k^\top \mathbf{R}_0). \tag{47}$$

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.7 presents a proof.

PROPOSITION 5. *Let $\|\mathbf{v}\|$ be the L^2 norm of vector $\mathbf{v}$. I assume that*

$$\max_{\|\mathbf{f}\|=1} \sigma(\tilde{M}(\mathbf{f})) \leq H \tag{48}$$

for some constant H. If $\sum_{n=K+1}^{N} \pi_n$ tends to zero, then $\mathbb{E}[\|\Delta\mathbf{p}(\mathbf{f}) - \Delta\check{\mathbf{p}}(\mathbf{f})\|^2]$ tends to zero uniformly for all F-SDF $\tilde{M}(\cdot)$.

Condition (48) requires the F-SDF $\tilde{M}(\mathbf{f})$ not to be overly volatile. Because the F-SDF $\tilde{M}(\mathbf{f})$ is linear in flow $\mathbf{f}$, I bound the F-SDF volatility on the sphere $\|\mathbf{f}\| = 1$. By the flow-based Hansen-Jagannathan bound, 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 (48) stipulates that these good deals should not exist.

The variance $\sum_{n=K+1}^{N} \pi_n$ of idiosyncratic flows tending to zero means that flows have a factor structure. Data support this empirical assumption (Hasbrouck and Seppi, 2001). 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 that I have a good approximation in the mean-squared-error sense. In sum, if the F-SDF is not overly volatile, the factor structure of flows implies the factor model of price impacts.

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 model (24) to some K^2 model (i.e., skipping from Proposition 3 to this section). Online Appendix D shows that the answer is no. Without the IUF, there is no proper theoretical justification for eliminating the cross-impacts of factor flows on portfolios that idiosyncratic flows go into.

4 Conclusion

I develop a new approach, model, and definition to study demand effects in asset pricing. My approach generalizes arbitrage pricing and avoids making any parametric assumptions on the utility function and payoff distribution, which are commonly found in equilibrium literature. By doing so, I reveal and relax an unrealistic cross-sectional restriction on price impacts present in the literature's quadratic-normal setup. In this process, I develop new theoretical tools that generalize textbook tools, including flow-based SDF, portfolio flow theory, flow-based Hansen-Jagannathan bound, IUF, and flow-based APT.

In my model, price impacts between underlying assets occur through factors that connect to these assets via the covariance structure of noisy flows and fundamental risks. Specifically, I develop a new definition for factor-level demand elasticity, highlighting that the conventional definition is inadequate and ill-defined.

References

Alekseev, Georgij, Stefano Giglio, Quinn Maingi, Julia Selgrad, and Johannes Stroebel, 2022, A quantity-based approach to constructing climate risk hedge portfolios, Working paper, NYU.

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, 2022, Flow-based asset pricing: A factor framework of cross-sectional price impacts, Working paper, Johns Hopkins University.

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

Basak, Suleyman, and Anna Pavlova, 2013, Asset prices and institutional investors, *American Economic Review* 103, 1728–1758.

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

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.

Chamberlain, Gary, 1983, Funds, factors, and diversification in arbitrage pricing models, *Econometrica* 1305–1323.

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.

Cremers, K. J. Martijn, and Jianping Mei, 2007, Turning over turnover, *Review of Financial Studies* 20, 1749–1782.

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.

Du, Wenxin, Alexander Tepper, and Adrien Verdelhan, 2018, Deviations from covered interest rate parity, *Journal of Finance* 73, 915–957.

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.

Frazzini, Andrea, and Owen A Lamont, 2008, Dumb money: Mutual fund flows and the cross-section of stock returns, *Journal of Financial Economics* 88, 299–322.

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

Gibbons, Michael R, Stephen A Ross, and Jay Shanken, 1989, A test of the efficiency of a given portfolio, *Econometrica* 1121–1152.

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

Hartzmark, Samuel M, and David H Solomon, 2022, Predictable price pressure, Working paper, University of Chicago.

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

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

Kodres, Laura E, and Matthew Pritsker, 2002, A rational expectations model of financial contagion, *Journal of Finance* 57, 769–799.

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.

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

Li, Jian, Zhiyu Fu, and Manav Chaudhary, 2022, Corporate bond elasticities: Substitutes matter, Working paper, Columbia University.

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.

Markowitz, Harry, 1952, Portfolio selection, *Journal of Finance* 7, 77–91.

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.

Raponi, Valentina, Raman Uppal, and Paolo Zaffaroni, 2022, Robust portfolio choice Working paper, IESE Business School.

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.

Shleifer, Andrei, and Robert W. Vishny, 1997, The limits of arbitrage, *Journal of Finance* 52, 35–55.

Vayanos, Dimitri, 2021, Price multipliers of anticipated and unanticipated shocks to demand and supply, Working paper, LSE.

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

Wurgler, Jeffrey, and Ekaterina Zhuravskaya, 2002, Does arbitrage flatten demand curves for stocks? *Journal of Business* 75, 583–608.

A Appendix for Proofs

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

A.1 Proof of Proposition 2

I write the F-SDF $\tilde{M}(\mathbf{f})$ in the fundamental-risk space spanned by $\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]$ as $\tilde{M}^*(\mathbf{f}) = (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])^\top \mathbf{b}$ for some $\mathbf{b} \in \mathbb{R}^N$. By equation (11), 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

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

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

I claim that

$$\mathrm{var}(\tilde{M}^*(\mathbf{f})) \leq \max_{\mathbf{c}\in\mathbb{R}^N} \frac{\Delta\mathbf{p}(\mathbf{f})^\top \mathbf{c}\mathbf{c}^\top \Delta\mathbf{p}(\mathbf{f})}{\mathrm{var}(\mathbf{c}^\top \mathbf{R}_0)}. \tag{A.2}$$

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

$$\frac{\Delta\mathbf{p}(\mathbf{f})^\top \mathbf{c}\mathbf{c}^\top \Delta\mathbf{p}(\mathbf{f})}{\mathrm{var}(\mathbf{c}^\top \mathbf{R}_0)} = \frac{(\Delta\mathbf{p}(\mathbf{f})^\top \mathrm{var}(\mathbf{R}_0)^{-1}\Delta\mathbf{p}(\mathbf{f}))^2}{\mathbf{c}^\top \mathrm{var}(\mathbf{R}_0)\mathbf{c}} = \Delta\mathbf{p}(\mathbf{f})^\top \mathrm{var}(\mathbf{R}_0)^{-1}\Delta\mathbf{p}(\mathbf{f}). \tag{A.3}$$

Therefore, I have proved (A.2). Combining with inequality (18) in the main text, I have proved the flow-based Hansen-Jagannathan bound.

A.2 Proof of Proposition 3

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

$$\begin{aligned}\tilde{M}(a_1\mathbf{f}_1 + a_2\mathbf{f}_2) =& \, \Delta\mathbf{p}(a_1\mathbf{f}_1 + a_2\mathbf{f}_2)^\top \mathrm{var}(\mathbf{R}_0)^{-1}(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) \\ =& \, (a_1\Delta\mathbf{p}(\mathbf{f}_1) + a_2\Delta\mathbf{p}(\mathbf{f}_2))^\top \mathrm{var}(\mathbf{R}_0)^{-1}(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) \\ =& \, a_1\tilde{M}(\mathbf{f}_1) + a_2\tilde{M}(\mathbf{f}_2). \end{aligned} \tag{A.4}$$

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.4) implies that $\tilde{M}(\mathbf{f}) = \sum_{n=1}^N f_n \tilde{M}(\mathbf{c}_n)$. Condition iii implies that $\tilde{M}(\mathbf{c}_n) \in \underline{\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]}$ for all $n = 1, 2, \ldots, N$. By Corollary 1, for each n, there exists a unique $\tilde{M}^*(\mathbf{c}_n) \in \underline{\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]}$, such that $\tilde{M}^*(\mathbf{c}_n) = \mathbf{y}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])$ for some $\mathbf{y}_n \in \mathbb{R}^N$. Therefore, I have

$$\tilde{M}(\mathbf{f}) = \sum_{n=1}^N f_n \mathbf{y}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) = (\mathbf{Y}\mathbf{f})^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]), \tag{A.5}$$

where $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$. Equation (24) and Assumption 2 imply that $\mathrm{var}(\mathbf{R}_0)\mathbf{Y}$ is positive definite.

A.3 Proof of Lemma 1

Because the matrix $\mathrm{var}(\mathbf{R}_0)$ has full rank, I carry out Cholesky decomposition and obtain $\mathrm{var}(\mathbf{R}_0) = \mathbf{U}^\top\mathbf{U}$, 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 $\mathbf{U}\mathrm{var}(\mathbf{f})\mathbf{U}^\top$. Because $\mathrm{var}(\mathbf{f})$ has full rank, I obtain

$$\mathbf{U}\mathrm{var}(\mathbf{f})\mathbf{U}^\top\mathbf{G} = \mathbf{G}\mathbf{\Pi}, \tag{A.6}$$

where $\mathbf{\Pi} = \mathrm{diag}(\pi_1, \pi_2, \ldots, \pi_N)$ 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 $\mathbf{B}$ satisfies the conditions (28) and (29). First, I have

$$\mathbf{B}^\top \mathrm{var}(\mathbf{R}_0)\mathbf{B} = \mathbf{G}^\top (\mathbf{U}^\top)^{-1}\mathbf{U}^\top \mathbf{U}\mathbf{U}^{-1}\mathbf{G} = \mathbf{I}_N. \tag{A.7}$$

Second, because $\mathbf{f} = \mathbf{B}\mathbf{q}$, I have

$$\mathrm{var}(\mathbf{f}) = \mathbf{B}\mathrm{var}(\mathbf{q})\mathbf{B}^\top. \tag{A.8}$$

From (A.6), I have $\mathbf{U}\mathrm{var}(\mathbf{f})\mathbf{U}^\top \mathbf{U}\mathbf{B} = \mathbf{U}\mathbf{B}\mathbf{\Pi}$. Because $\mathbf{U}$ is invertible, I have

$$\mathrm{var}(\mathbf{f})\mathbf{U}^\top \mathbf{U}\mathbf{B} = \mathbf{B}\mathbf{\Pi}. \tag{A.9}$$

Plugging (A.8) into (A.9), I obtain

$$\mathbf{B}\mathbf{\Pi} = \mathbf{B}\mathrm{var}(\mathbf{q})\mathbf{B}^\top \mathbf{U}^\top \mathbf{U}\mathbf{B} = \mathbf{B}\mathrm{var}(\mathbf{q}), \tag{A.10}$$

where I have used $\mathbf{B}^\top \mathbf{U}^\top \mathbf{U}\mathbf{B} = \mathbf{G}^\top \mathbf{G} = \mathbf{I}_N$. Equation (A.10) implies $\mathrm{var}(\mathbf{q}) = \mathbf{\Pi}$.

Finally, I show the uniqueness of the matrix $\mathbf{B}$. Suppose that some other matrix $\tilde{\mathbf{B}}$ satisfies (28) and (29). I define $\tilde{\mathbf{G}} = \mathbf{U}\tilde{\mathbf{B}}$. I then have $\tilde{\mathbf{G}}^\top \tilde{\mathbf{G}} = \mathbf{I}_N$ and $\tilde{\mathbf{G}}^\top \mathbf{U}\mathrm{var}(\mathbf{f})\mathbf{U}^\top \tilde{\mathbf{G}} = \mathbf{\Pi}$. Assumption 4 requires that $\mathbf{U}\mathrm{var}(\mathbf{f})\mathbf{U}^\top$ has distinct eigenvalues, which implies that matrix $\tilde{\mathbf{G}}$ is unique up to multiplication by -1 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 multiplication by -1 for any columns.

A.4 Proof of Theorem 1

I start by showing the following lemma, which allows me to state the equivalent IUF condition under the orthogonalized 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* $\mathrm{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{a}^\top \mathbf{R}_0)\mathrm{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) = 0$ *for all* $n = 1, 2, \ldots, N$.

Proof. Following the proof of Lemma 1 in Appendix A.3, I carry out Cholesky decomposition to obtain $\mathrm{var}(\mathbf{R}_0) = \mathbf{U}^\top \mathbf{U}$, 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 $\mathrm{var}(\mathbf{Uf})$ with distinct eigenvalues $\pi_1 > \pi_2 > \cdots > \pi_N > 0$.

I have that

$$
\begin{aligned}
& \mathrm{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{a}^\top \mathbf{R}_0)\mathrm{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) = 0 \text{ for all } n = 1, 2, \ldots, N \\
\Longleftrightarrow\ & \mathbf{a}^\top \mathrm{var}(\mathbf{R}_0)\mathbf{b}_n \mathbf{b}_n^\top \mathrm{var}(\mathbf{R}_0)\mathbf{d} = 0 \text{ for all } n = 1, 2, \ldots, N \\
\Longleftrightarrow\ & \mathbf{a}^\top \mathbf{U}^\top \mathbf{g}_n \mathbf{g}_n^\top \mathbf{U}\mathbf{d} = 0 \text{ for all } n = 1, 2, \ldots, N && \text{(A.11)} \\
\Longleftrightarrow\ & \mathbf{a}^\top \mathbf{U}^\top \left(\sum_{n=1}^{N} z_n \mathbf{g}_n \mathbf{g}_n^\top \right) \mathbf{U}\mathbf{d} = 0 \text{ for any real numbers } z_1, z_2, \ldots, z_N, && \text{(A.12)}
\end{aligned}
$$

where I have used the Cholesky decomposition of $\mathrm{var}(\mathbf{R}_0)$ in (A.11).

I define the following set of matrices,

$$
\mathcal{H} = \left\{ \mathbf{H} \in \mathbb{R}^{N \times N} \,\middle|\, \mathbf{H} = \sum_{n=1}^{N} z_n \mathbf{g}_n \mathbf{g}_n^\top \text{ for some real numbers } z_1, z_2, \ldots, z_N \right\}. \qquad \text{(A.13)}
$$

I claim that

$$
\mathcal{H} = \left\{ \mathbf{H} \in \mathbb{R}^{N \times N} \,\middle|\, \mathbf{H}\mathbf{U}\mathrm{var}(\mathbf{f})\mathbf{U}^\top = \mathbf{U}\mathrm{var}(\mathbf{f})\mathbf{U}^\top \mathbf{H} \right\}. \qquad \text{(A.14)}
$$

It is easy to see that if $\mathbf{H} \in \mathcal{H}$ as defined in (A.13), I have $\mathbf{H} \in \mathcal{H}$ as defined in (A.14) because $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N$ are the orthogonalized eigenvectors of $\mathbf{U}\mathrm{var}(\mathbf{f})\mathbf{U}^\top$. Conversely, if $\mathbf{H} \in \mathcal{H}$ as defined in (A.14), then for any n,

$$\mathbf{U}\mathrm{var}(\mathbf{f})\mathbf{U}^\top(\mathbf{H}\mathbf{g}_n) = (\mathbf{U}\mathrm{var}(\mathbf{f})\mathbf{U}^\top\mathbf{H})\mathbf{g}_n = (\mathbf{H}\mathbf{U}\mathrm{var}(\mathbf{f})\mathbf{U}^\top)\mathbf{g}_n = \pi_n(\mathbf{H}\mathbf{g}_n), \tag{A.15}$$

showing that $\mathbf{H}\mathbf{g}_n$ is also the eigenvector of $\mathbf{U}\mathrm{var}(\mathbf{f})\mathbf{U}^\top$ 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.13).

Given (A.14), I can equivalently rewrite (A.12) as

$$\mathbf{a}^\top\mathbf{U}^\top\mathbf{H}\mathbf{U}\mathbf{d} = 0 \text{ for any } \mathbf{H} \in \mathcal{H}. \tag{A.16}$$

I define $\mathbf{C} = \mathbf{U}^\top\mathbf{H}\mathbf{U}$. Using the fact that $\mathbf{U}$ is invertible and $\mathrm{var}(\mathbf{R}_0) = \mathbf{U}^\top\mathbf{U}$, I have $\mathbf{C} \in \mathcal{C} \iff \mathbf{H} \in \mathcal{H}$, where $\mathcal{C}$ is defined in (25). Therefore, I can equivalently rewrite (A.16) 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 orthogonalized $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$.

Assumption 3*. ***IUF under orthogonalized portfolios***

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

$$\mathbf{d} = (\mathrm{cov}(f_1, s), \mathrm{cov}(f_2, s), \ldots, \mathrm{cov}(f_N, s))^\top. \tag{A.17}$$

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

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

I also prove the following coordinate transformation,

$$
\begin{aligned}
\mathbf{d} &= (\mathrm{cov}(f_1, s), \mathrm{cov}(f_2, s), \dots, \mathrm{cov}(f_N, s))^\top && \text{(A.18)}\\
&= \left(\mathrm{cov}\left(\sum_{n=1}^{N} b_{1,n} q_n, s\right), \mathrm{cov}\left(\sum_{n=1}^{N} b_{2,n} q_n, s\right), \dots, \mathrm{cov}\left(\sum_{n=1}^{N} b_{N,n} q_n, s\right)\right)^\top && \text{(A.19)}\\
&= \left(\sum_{n=1}^{N} b_{1,n}\mathrm{cov}(q_n, s), \sum_{n=1}^{N} b_{2,n}\mathrm{cov}(q_n, s), \dots, \sum_{n=1}^{N} b_{N,n}\mathrm{cov}(q_n, s)\right)^\top && \text{(A.20)}\\
&= \mathbf{B}\,(\mathrm{cov}(q_1, s), \mathrm{cov}(q_2, s), \dots, \mathrm{cov}(q_N, s))^\top && \text{(A.21)}\\
&= \sum_{n=1}^{N} \mathbf{b}_n \mathrm{cov}(q_n, s), && \text{(A.22)}
\end{aligned}
$$

where equation (A.18) follows from definition (26), equation (A.19) uses definition (27), and $b_{n,k}$ is the (n,k)-th element of matrix $\mathbf{B}$. Equations (A.20) and (A.21) are straightforward algebraic manipulations, and equation (A.22) follows from $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_N)$. This coordinate transformation (A.22) 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.22) and orthogonalization condition $\mathbf{B}^\top \mathrm{var}(\mathbf{R}_0)\mathbf{B} = \mathbf{I}_N$ in (28), I have

$$
\mathrm{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) = \mathbf{b}_n^\top \mathrm{var}(\mathbf{R}_0)\left(\sum_{m=1}^{N} \mathbf{b}_m \mathrm{cov}(q_m, s)\right) = \mathrm{cov}(q_n, s). \tag{A.23}
$$

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 canonical form (30) satisfies the IUF in Assumption 3*. I derive the price impact under the canonical form (30) as

$$
\Delta \mathbf{p}(\mathbf{f}) = \mathbb{E}[\mathbf{R}_0 \tilde{M}(\mathbf{f})] = \mathrm{var}(\mathbf{R}_0)\mathbf{B}\boldsymbol{\Lambda}\mathbf{q} = \mathrm{var}(\mathbf{R}_0)\sum_{n=1}^{N} \lambda_n q_n \mathbf{b}_n. \tag{A.24}
$$

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

$$\mathbf{a} = \sum_{n=1}^{N} x_n \mathbf{b}_n \tag{A.25}$$

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

$$\begin{aligned} 0 &= \mathrm{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{a}^\top \mathbf{R}_0)\mathrm{cov}(\mathbf{b}_n^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) \\ &= \left(\sum_{m=1}^{N} x_m \mathbf{b}_m\right)^\top \mathrm{var}(\mathbf{R}_0)\mathbf{b}_n \mathrm{cov}(q_n, s) = x_n \mathrm{cov}(q_n, s), \end{aligned} \tag{A.26}$$

for any $n = 1, 2, \dots, N$, where I use equations (28), (A.23), and (A.25).

I have

$$\mathbf{a}^\top \Delta \mathbf{p}(\mathbf{f}) = \left(\sum_{m=1}^{N} x_m \mathbf{b}_m\right)^\top \mathrm{var}(\mathbf{R}_0)\left(\sum_{n=1}^{N} \lambda_n q_n \mathbf{b}_n\right) = \sum_{n=1}^{N} x_n \lambda_n q_n, \tag{A.27}$$

where I use equations (28), (A.24), and (A.25). Therefore, I have

$$\mathrm{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\mathbf{f}), s) = \sum_{n=1}^{N} x_n \mathrm{cov}(q_n, s)\lambda_n. \tag{A.28}$$

Note that the IUF's assumption implies that $x_n \mathrm{cov}(q_n, s) = 0$ for any $n = 1, 2, \dots, N$ in (A.26). Therefore, by equation (A.28), I have $\mathrm{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\mathbf{f}), s) = 0$, thereby showing that the IUF holds. Moreover, equation (35) implies that

$$\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f}) = \mathbf{q}^\top \mathbf{B}^\top \mathrm{var}(\mathbf{R}_0)\mathbf{B}\boldsymbol{\Lambda}\mathbf{q} = \mathbf{q}^\top \boldsymbol{\Lambda}\mathbf{q}. \tag{A.29}$$

Because $\boldsymbol{\Lambda}$ is positive definite, Assumption 2 holds.

Second, I show that, if the unique F-SDF in (23) satisfies the IUF, the F-SDF can be written in the canonical form. Using equation (27), I simplify (23) as

$$\begin{aligned}\tilde{M}(\mathbf{f}) &= \sum_{n=1}^{N}\sum_{m=1}^{N} q_m b_{n,m}\mathbf{g}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) \\ &= \sum_{m=1}^{N} q_m \sum_{n=1}^{N} b_{n,m}\mathbf{g}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) = \sum_{m=1}^{N} q_m \mathbf{h}_m^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]), \qquad \text{(A.30)}\end{aligned}$$

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.30) is

$$\Delta\mathbf{p}(\mathbf{f}) = \mathbb{E}[\mathbf{R}_0\tilde{M}(\mathbf{f})] = \mathrm{var}(\mathbf{R}_0)\sum_{m=1}^{N} q_m\mathbf{h}_m. \qquad \text{(A.31)}$$

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.22),

$$\mathbf{d} = \sum_{n=1}^{N} \mathbf{b}_n \mathrm{cov}(q_n, q_m) = \pi_m \mathbf{b}_m, \qquad \text{(A.32)}$$

because $\mathrm{var}(\mathbf{q}) = \mathrm{diag}(\pi_1, \pi_2, \ldots, \pi_N)$. Thus, I have $\mathrm{cov}(\mathbf{b}_n^\top\mathbf{R}_0, \mathbf{a}^\top\mathbf{R}_0)\mathrm{cov}(\mathbf{b}_n^\top\mathbf{R}_0, \mathbf{d}^\top\mathbf{R}_0) = 0$ for any $n = 1, 2, \ldots, N$, because $\mathbf{B}^\top\mathrm{var}(\mathbf{R}_0)\mathbf{B} = \mathbf{I}_N$. Thus, the IUF implies that

$$0 = \mathrm{cov}(\mathbf{a}^\top\Delta\mathbf{p}(\mathbf{f}), s) = \mathrm{cov}\left(\mathbf{b}_l^\top\mathrm{var}(\mathbf{R}_0)\sum_{n=1}^{N} q_n\mathbf{h}_n, q_m\right) = \pi_m\mathbf{b}_l^\top\mathrm{var}(\mathbf{R}_0)\mathbf{h}_m. \qquad \text{(A.33)}$$

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

$$\mathbf{b}_l^\top\mathrm{var}(\mathbf{R}_0)\mathbf{h}_m = 0. \qquad \text{(A.34)}$$

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.34) and $\mathbf{B}^\top\mathrm{var}(\mathbf{R}_0)\mathbf{B} = \mathbf{I}_N$, I have $\theta_{m,n} = 0$ for any $m \neq n$. Therefore, I can rewrite the F-SDF form (A.30) as

$$\tilde{M}(\mathbf{f}) = \sum_{n=1}^{N}\theta_{n,n}q_n\mathbf{b}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]). \qquad \text{(A.35)}$$

I rename $\theta_{n,n} = \lambda_n$ and obtain the canonical form (30). Lastly, using (A.29) and Assumption 2, I have $\lambda_n > 0$ for all n. The proof is complete.

A.5 Proof of Proposition 4

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 $\mathbf{B}$ and $\mathbf{q}$ are the orthogonalized factor portfolios and flows in the canonical form (30). By (31), I have

$$\tilde{M}(\mathbf{f}) = (\mathbf{O}^{-1}\tilde{\mathbf{\Lambda}}\mathbf{O}\tilde{\mathbf{q}})^\top \mathbf{B}^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]). \tag{A.36}$$

By (30), the above equation implies that $\mathbf{O}^{-1}\tilde{\mathbf{\Lambda}}\mathbf{O} = \mathbf{\Lambda}$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ is the diagonal price-of-flow-induced-risk matrix. Because the orthogonalized $\mathbf{B}$ and $\mathbf{q}$ satisfy Lemma 1, conditions (28) and (29) translate into (33) and (34).

A.6 Proof of Theorem 2

I first show the sufficiency direction. From Theorem 1 and Proposition 4, if the orthogonalized model (42) 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. Because $\tilde{\mathbf{B}}\mathbf{O} = \mathbf{B}$ and $\mathbf{O}\mathbf{q} = \tilde{\mathbf{q}}$, 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 (46), note that I have, by (41) and $\text{cov}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0) = 0$ for $n \neq l$,

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

I then have

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

Similarly, I have, by (42),

$$\tilde{\mathbf{b}}_l^\top \Delta \bar{\mathbf{p}}(\mathbf{f}) = \sum_{n=1}^N \tilde{\lambda}_{n,n} \tilde{q}_n \mathrm{cov}(\tilde{\mathbf{b}}_l \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0) = \tilde{\lambda}_{l,l} \tilde{q}_l \mathrm{var}(\tilde{\mathbf{b}}_l \mathbf{R}_0), \tag{A.39}$$

and

$$\frac{\partial \tilde{\mathbf{b}}_l^\top \Delta \bar{\mathbf{p}}(\mathbf{f})}{\partial \tilde{q}_l} = \tilde{\lambda}_{l,l} \mathrm{var}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0). \tag{A.40}$$

This shows that (46) holds.

Under the price impact model (41), the expected compensation for risk is

$$\begin{aligned}\mathbb{E}[\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f})] =& \mathbb{E}[\tilde{\mathbf{q}}^\top \tilde{\mathbf{B}}^\top \mathrm{var}(\mathbf{R}_0) \tilde{\mathbf{B}} \tilde{\mathbf{\Lambda}} \tilde{\mathbf{q}}] = \sum_{n=1}^N \sum_{m=1}^N \tilde{\lambda}_{n,m} \mathrm{var}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0) \mathrm{cov}(\tilde{q}_m, \tilde{q}_n) &\text{(A.41)}\\ =& \sum_{n=1}^N \tilde{\lambda}_{n,n} \mathrm{var}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0) \mathrm{var}(\tilde{q}_n), &\text{(A.42)}\end{aligned}$$

where the second equality in (A.41) uses $\mathrm{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ and equality (A.42) uses $\mathrm{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ for any $n \neq m$. Therefore, the expected compensation $\mathbb{E}[\mathbf{f}^\top \Delta \mathbf{p}(\mathbf{f})]$ under models (41) and (42) are the same for any N^2 parameters $\tilde{\lambda}_{n,m}$. By equation (45), the expected utility $\mathbb{E}[u(W(\mathbf{f}))]$ under models (41) and (42) are also the same.

Next, I show the necessity direction. I have, by (41),

$$\tilde{\mathbf{b}}_l^\top \Delta \mathbf{p}(\mathbf{f}) = \sum_{n=1}^N \sum_{m=1}^N \tilde{\lambda}_{n,m} \tilde{q}_m \mathrm{cov}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0). \tag{A.43}$$

I then have

$$\frac{\partial \tilde{\mathbf{b}}_l^\top \Delta \mathbf{p}(\mathbf{f})}{\partial \tilde{q}_l} = \sum_{n=1}^N \tilde{\lambda}_{n,l} \mathrm{cov}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0). \tag{A.44}$$

Similarly, I have, by (42),

$$\tilde{\mathbf{b}}_l^\top \Delta\bar{\mathbf{p}}(\mathbf{f}) = \sum_{n=1}^{N} \tilde{\lambda}_{n,n} \tilde{q}_n \mathrm{cov}(\tilde{\mathbf{b}}_l \mathbf{R}_0, \tilde{\mathbf{b}}_n^\top \mathbf{R}_0), \tag{A.45}$$

and

$$\frac{\partial \tilde{\mathbf{b}}_l^\top \Delta\check{\mathbf{p}}(\mathbf{f})}{\partial \tilde{q}_l} = \tilde{\lambda}_{l,l} \mathrm{var}(\tilde{\mathbf{b}}_l^\top \mathbf{R}_0). \tag{A.46}$$

Because (46) holds for any $\tilde{\mathbf{\Lambda}} = \{\tilde{\lambda}_{n,m}\}$, $\mathrm{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ for any $n \neq m$.

Under the price impact model (41), the expected compensation is

$$\mathbb{E}[\mathbf{f}^\top \Delta\mathbf{p}(\mathbf{f})] = \mathbb{E}[\tilde{\mathbf{q}}^\top \tilde{\mathbf{B}}^\top \mathrm{var}(\mathbf{R}_0) \tilde{\mathbf{B}} \tilde{\mathbf{\Lambda}} \tilde{\mathbf{q}}] = \sum_{n=1}^{N} \sum_{m=1}^{N} \tilde{\lambda}_{n,m} \mathrm{var}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0) \mathrm{cov}\left(\tilde{q}_n, \tilde{q}_m\right), \tag{A.47}$$

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 (45), if the expected utility $\mathbb{E}[u(W(\mathbf{f}))]$ remains invariant under models (41) and (42), 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 (42) satisfies the IUF for any parameters $\tilde{\lambda}_{1,1}, \tilde{\lambda}_{2,2}, \ldots, \tilde{\lambda}_{N,N}$.

A.7 Proof of Proposition 5

By the flow-based Hansen-Jagannathan bound in Proposition 2, condition (48) implies

$$\max_{\|\mathbf{f}\|=1} \Delta\mathbf{p}(\mathbf{f})^\top \mathrm{var}(\mathbf{R}_0)^{-1} \Delta\mathbf{p}(\mathbf{f}) \leq H^2. \tag{A.48}$$

Under the N-factor model (35), the squared maximum price impact ratio is

$$\Delta\mathbf{p}(\mathbf{f})^\top \mathrm{var}(\mathbf{R}_0)^{-1} \Delta\mathbf{p}(\mathbf{f}) = \sum_{n=1}^{N} \lambda_n q_n \mathbf{b}_n^\top \mathrm{var}(\mathbf{R}_0) \sum_{n=1}^{N} \lambda_n q_n \mathbf{b}_n = \sum_{n=1}^{N} \lambda_n^2 q_n^2, \tag{A.49}$$

where I use the orthogonalization $\mathbf{B}^\top \mathrm{var}(\mathbf{R}_0)\mathbf{B} = \mathbf{I}_N$. 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.48) and (A.49), I have $|\lambda_m| \leq H\|\mathbf{b}_m\|$. That is, all price-of-flow-induced-risk coefficients λ_n are uniformly bounded for all F-SDF $\tilde{M}(\cdot)$.

I can simplify the pricing error of the price impact as

$$\Delta\mathbf{p}(\mathbf{f}) - \Delta\bar{\mathbf{p}}(\mathbf{f}) = \mathrm{var}(\mathbf{R}_0) \sum_{n=K+1}^{N} \lambda_n q_n \mathbf{b}_n = \sum_{n=K+1}^{N} \mathbf{c}_n q_n, \tag{A.50}$$

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

$$\|\Delta\mathbf{p}(\mathbf{f}) - \Delta\check{\mathbf{p}}(\mathbf{f})\|^2 = \sum_{n=1}^{N} \left(\sum_{k=K+1}^{N} c_{n,k} q_k \right)^2. \tag{A.51}$$

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

$$\mathbb{E}[\|\Delta\mathbf{p}(\mathbf{f}) - \Delta\check{\mathbf{p}}(\mathbf{f})\|^2] = \sum_{n=1}^{N} \mathbb{E} \left(\sum_{k=K+1}^{N} c_{n,k} q_k \right)^2 = \sum_{n=1}^{N} \sum_{k=K+1}^{N} c_{n,k}^2 \pi_k. \tag{A.52}$$

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

Online Appendix of
“Flow-Based Arbitrage Pricing Theory”

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

A General Theory of F-SDF with Duplicate Eigenvalues

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

A.1 Linear Model of the F-SDF

To state the canonical form of the 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 orthogonalized factors $\mathbf{B} = \mathbf{U}^{-1}\mathbf{G}$ following the proof of Lemma 1 in the main text.

DEFINITION O.1. *The canonical form of the F-SDF is*

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

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* $j = 1, 2, \ldots, J$.

I now prove Theorem 1 in the general case. First, I show that the canonical form (OA.3) satisfies the IUF. The price impact under the canonical form (OA.3) is

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

For any portfolio $\mathbf{a} \in \mathbb{R}^N$, flow s, and the corresponding portfolio $\mathbf{d}$ that satisfy $\mathbf{a}^\top \mathbf{C}\mathbf{d} = 0$ for all $\mathbf{C} \in \mathcal{C}$, the goal is to show $\mathrm{cov}(\mathbf{a}^\top \Delta\mathbf{p}(\mathbf{f}), s) = 0$.

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

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

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

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

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

$x_l \mathrm{cov}(q_m, s) = 0$ 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.22) in the main text and (OA.6), I simplify the IUF condition as

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

I define the set

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

Recall that I define the matrix $\mathbf{B} = \mathbf{U}^{-1}\mathbf{G}$ from a given of set of eigenvectors $\mathbf{G}$ of matrix $\mathbf{U}\mathrm{var}(\mathbf{f})\mathbf{U}^\top$ (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 $\mathbf{x}^\top \mathbf{G}^\top \mathbf{H}\mathbf{G}\mathrm{cov}(\mathbf{q}, s) = 0$ 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

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

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

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

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

Moreover, equation (OA.5) implies that

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

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

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

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

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

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.22) in the main text, I have

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

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

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

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

which simplifies to

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

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

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

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 $\mathbf{\Pi v} = \pi_j \mathbf{v}$. Therefore, I have

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

Therefore, $\mathbf{Lv}$ 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

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

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. The proof is complete.

Lastly, I present the general linear model of the F-SDF when $\text{var}(\mathbf{Uf})$ have duplicate eigenvalues. It is straightforward to see that Proposition 4 in the main text still holds, except that $\mathbf{\Lambda}$ is replaced by the block-diagonal form (OA.4).

A.2 Optimality of the IUF Orthogonalization

I show the optimality of the IUF orthogonalization, without imposing regularity Assumption 4 in the main text.

In the general case, Definition 2 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.2. *Each set of N linearly independent portfolios $\tilde{\mathbf{B}} = (\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N)$ defines a model orthogonalization. Specifically, the N^2 model 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 $\tilde{\mathbf{\Psi}}_{j,l}$ is an $r_j \times r_l$ matrix. The orthogonalized N-factor model under portfolios $\tilde{\mathbf{B}}$ is *defined as*

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

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

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

I now generalize Theorem 2 of the main text.

THEOREM O.1. *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 $\tilde{\boldsymbol{\Psi}}_{1,1}, \tilde{\boldsymbol{\Psi}}_{2,2}, \ldots, \tilde{\boldsymbol{\Psi}}_{N,N}$ if and only if for any N^2 parameters $\tilde{\boldsymbol{\Lambda}}$ 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 $\tilde{\boldsymbol{\Psi}}_{1,1}, \tilde{\boldsymbol{\Psi}}_{2,2}, \ldots, \tilde{\boldsymbol{\Psi}}_{N,N}$ if and only if $\mathrm{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ and $\mathrm{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ whenever the (n, n)-th and (m, m)-th elements of matrix $\boldsymbol{\Pi}$ correspond to distinct eigenvalues.

Proof. First, I show the sufficiency direction. From Theorem 1 of the general case, 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 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 $\tilde{\mathbf{\Gamma}}_{j,j}$ 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 orthogonalized factor portfolios in canonical form. Therefore, $\text{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ and $\text{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ whenever (n,n)-th and (m,m)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues.

Next, I show the necessity direction. Because $\text{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ 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, all I need for the canonical 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 canonical form (OA.4) satisfies the IUF by Theorem 1. □

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

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

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

I then have

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

Similarly, I have by (OA.26),

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

and

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

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 $\mathrm{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ 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 $\mathrm{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ 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 $\mathrm{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ whenever the (n, n)-th and (m, m)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues.

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

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

where each $\tilde{\mathbf{\Phi}}_{j,j}$ is an $r_j \times r_j$ matrix and I use $\mathrm{cov}(\tilde{\mathbf{b}}_n^\top \mathbf{R}_0, \tilde{\mathbf{b}}_m^\top \mathbf{R}_0) = 0$ 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

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

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

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

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 $\mathrm{cov}(\tilde{\mathbf{q}}_j, \tilde{\mathbf{q}}_l) = \mathbf{0}$ for any $j \neq l$. The last statement is precisely $\mathrm{cov}(\tilde{q}_n, \tilde{q}_m) = 0$ whenever the (n, n)-th and (m, m)-th elements of matrix $\mathbf{\Pi}$

correspond to distinct eigenvalues. □

The three claims together imply that Theorem O.1 is true. □

B Portfolio Construction Via Weighted Least Squares

In this appendix, I consider portfolio construction via weighted least squares (WLS).

As in the main text, I denote the N-asset flows as $\mathbf{f}$ and the K portfolios' weights as $\mathbf{B}$. I consider the portfolio flows constructed via WLS as $\mathbf{q} = (\mathbf{B}^\top \mathbf{W} \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{W} \mathbf{f}$ for some $N \times N$ weighting matrix $\mathbf{W}$.

To understand the WLS construction, I conduct Cholesky decomposition to obtain $\mathbf{W} = (\mathbf{U}^{-1})^\top \mathbf{U}^{-1}$, for some $N \times N$ upper triangular matrix $\mathbf{U}$. I consider the N portfolios given by the portfolio-weight matrix $\mathbf{U}$. Following Section 2.2 in the main text, flows into these N portfolios are $\tilde{\mathbf{f}} = \mathbf{U}^{-1}\mathbf{f}$. Note that both $\mathbf{U}$ and $\mathbf{B}$ are portfolio weights in terms of the N underlying assets. I now consider the portfolio weights $\tilde{\mathbf{B}}$ of the K portfolios (with weights $\mathbf{B}$ on the N assets) in terms of the N portfolios (with weights $\mathbf{U}$ on the N assets). By definition, $b_{n,k} = \sum_{m=1}^{N} u_{n,m} \tilde{b}_{m,k}$, where $b_{n,k}$ is the dollar amount of asset n held by portfolio k (with weights $\mathbf{B}$), $u_{n,m}$ is the dollar amount of asset n held by portfolio m (with weights $\mathbf{U}$), and $\tilde{b}_{m,k}$ is the dollar amount of portfolio m (with weights $\mathbf{U}$) held by portfolio k (with weights $\mathbf{B}$). By matrix multiplication rule, I have $\mathbf{B} = \mathbf{U}\tilde{\mathbf{B}}$.

With $\tilde{\mathbf{f}} = \mathbf{U}^{-1}\mathbf{f}$ and $\mathbf{B} = \mathbf{U}\tilde{\mathbf{B}}$, I have

$$\mathbf{q} = (\mathbf{B}^\top \mathbf{W} \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{W} \mathbf{f} = (\tilde{\mathbf{B}}^\top \tilde{\mathbf{B}})^{-1} \tilde{\mathbf{B}}^\top \tilde{\mathbf{f}}. \tag{OB.1}$$

Therefore, the WLS construction can be interpreted as first rotating the N-asset flows $\mathbf{f}$ into N equivalent portfolios $\mathbf{U}$ and then projecting these N portfolios onto the K portfolio weights $\tilde{\mathbf{B}}$, which is the equivalent of $\mathbf{B}$ but is expressed in terms of the N portfolios.

C Static CARA-normal Price Impact Model

In this appendix, I show that the static CARA-normal price impact model is a special case of my N-factor price impact model 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, and I write the flow as an $N \times 1$ vector as $\mathbf{h} = (h_1, h_2, \ldots, h_N)$. Unlike in the baseline setup, the unit of flow is expressed in the number of shares, not dollar amounts. I later recover the flow in dollar amounts. 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 2.1. I now introduce the CARA-normal assumptions:

1. Each arbitrageur has 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

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

where the time-1 wealth of each arbitrageur is

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

Standard calculation implies that

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

Therefore, the price change is

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

I define price impact as

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

and fundamental return as

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

Using the fact that

$$\mathrm{var}(\mathbf{X}) = \mathrm{diag}(P_1(\mathbf{0}), P_2(\mathbf{0}), \dots, P_N(\mathbf{0}))\mathrm{var}(\mathbf{R}_0)\mathrm{diag}(P_1(\mathbf{0}), P_2(\mathbf{0}), \dots, P_N(\mathbf{0})), \tag{OC.7}$$

I have

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

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

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

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

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

Using the orthogonalized factor portfolios and flows from Lemma 1 in the main text, I have $\mathbf{f} = \sum_{n=1}^{N} q_n \mathbf{b}_n$. Therefore, the CARA-normal price impact model is

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

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

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

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

D 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

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

and aim to approximate this model by the first K factors

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

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?

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

		factor flow $q_1, q_2, \ldots, q_K$	idiosyncratic flow $q_{K+1}, q_{K+2}, \ldots, q_N$
factor portfolio	$\mathbf{b}_1$ $\mathbf{b}_2$ $\vdots$ $\mathbf{b}_K$	$K \times K$ price of flow-induced risk $\lambda_{n,m}$	$K \times (N-K)$ price of flow-induced risk $\lambda_{n,m}$ pricing error tends to zero
idiosyncratic portfolio	$\mathbf{b}_{K+1}$ $\mathbf{b}_{K+2}$ $\vdots$ $\mathbf{b}_N$	$(N-K) \times K$ price of flow-induced risk $\lambda_{n,m}$ pricing error does NOT tend to zero	$(N-K) \times (N-K)$ price of flow-induced risk $\lambda_{n,m}$ pricing error tends to zero

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, as 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. My IUF first orthogonalizes the N^2 prices of flow-induced risk to the N diagonal terms using the commonality in flow-induced risk. The flow-based APT then reduces N diagonal prices of flow-induced risk to K using the factor structure of flow-induced risk.

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

However, note that

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

The second term is the price impact of factor flows on portfolios that idiosyncratic flows go into, which does not tend 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 $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N$ using the commonality in flow-induced risk. Intuitively, idiosyncratic portfolios have both negligible flows and negligible risks. The IUF implies that these orthogonalized portfolios have no cross impacts. In Figure O.1, the IUF first orthogonalizes the N^2 prices of flow-induced risk to the N diagonal terms. The flow-based APT then reduces the N diagonal prices of flow-induced risk to K using the factor structure of flow-induced risk.