An Arbitrage Foundation for Demand Effects in Asset Pricing*

Yu An

July 30, 2025

Abstract

Demand for any of N assets can influence every other price, generating N^2 cross-impact slopes. Which interactions are determined by economic first principles and which reflect investor behavior? Using no arbitrage alone—without needing to specify a particular equilibrium—a new "Irrelevance of Uncorrelated Portfolios" axiom is necessary and sufficient to isolate N factor portfolios whose demand moves only its own price of risk, collapsing the description to N factor-specific slopes. Two further restrictions yield contrasting simplifications: equal slopes reproduce the familiar mean-variance formula, whereas a demand-based APT compresses the model to a few factor slopes plus one idiosyncratic elasticity.

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

JEL Codes: G11, G12

*Johns Hopkins Carey Business School. Email: yua@jhu.edu. I especially thank Yinan Su and Chen Wang for numerous conversations. I thank Yun Joo An, Federico Bandi, Darrell Duffie, Nicola Fusari, Amy Huber, Evan Jo, Praveen Kumar, Mina Lee, Jane Li, Jiacui Li, Wenhao Li, Daniel Neuhann, Nagpurnanand Prabhala, Marzena Rostek, Alessandro Rebucci, Shrihari Santosh, Zhaogang Song, Semih Üslü, Milena Wittwer, Zeyu Zheng, and seminar participants at JHU Carey, USC finance reading group, Australasian finance & banking conference, MFA, University of Colorado Boulder, Eastern Finance Association, FMCG, and NFA for helpful comments. This paper was previously circulated under the title "Flow-Based Arbitrage Pricing Theory." All errors are my own.

1 Introduction

Uninformed trading demand can push individual asset prices away from fundamental values, whereas common risk exposures pull prices together. Empirically, both forces are strong: demand shocks have large, immediate effects on prices (Gabaix and Koijen, 2022), yet a handful of risk factors account for most price co-movement (Fama and French, 1993). Their coexistence raises a natural question. With N assets, a shock to the demand for one asset can, in principle, influence the prices of the other $N-1$ assets. The reduced-form description therefore contains N^2 cross-impact slopes that record "who moves whom." Which of those N^2 interactions follow directly from economic first principles, and which ones are free to reflect investor behavior?

Much of the existing literature answers by positing a specific equilibrium model—either a demand-system model à la Koijen and Yogo (2019) or a mean-variance framework à la Vayanos and Vila (2021). Those frameworks produce tractable cross-impact formulas, but only after imposing particular utility functions, payoff distributions, or market frictions—features that are difficult to observe and verify empirically. When a restriction appears in the final formula, it can be difficult to know how much stems from fundamental economic forces and how much reflects the particular equilibrium machinery (Fuchs, Fukuda, and Neuhann, 2025a,b; Koijen and Yogo, 2025).

This paper reverses the usual order of reasoning. It starts with the most general no-arbitrage price-impact representation—an N^2-parameter object—and then imposes a sequence of economic restrictions that progressively reduce dimensionality. First, the *Irrelevance of Uncorrelated Portfolios* (IUP) axiom collapses the N^2 system to an N-parameter factor model in which demand for each factor moves only that factor's own price of risk.

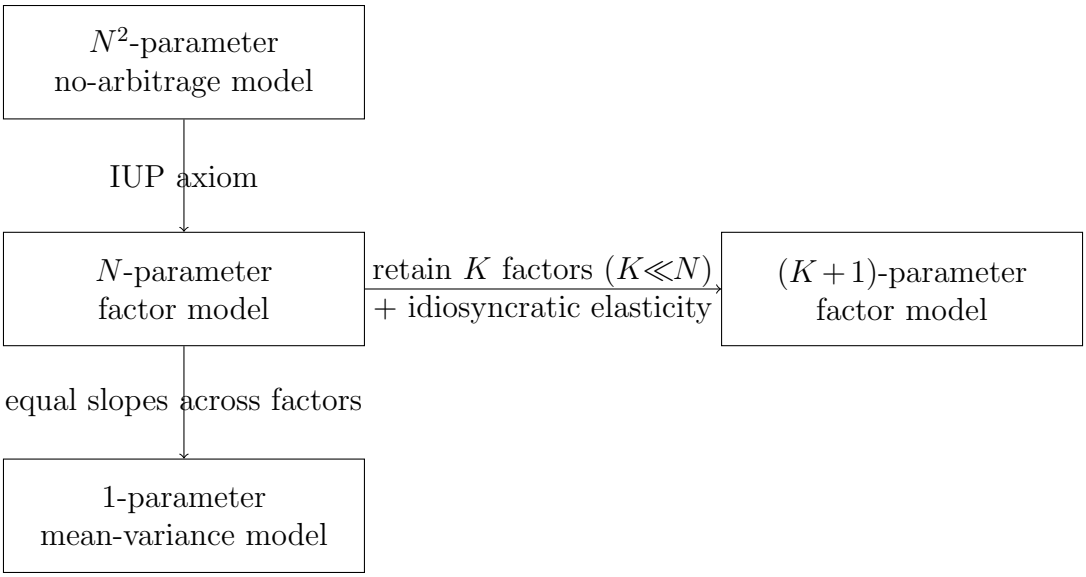


Figure 1. Dimensionality reductions produced by successive economic conditions.

Adding an *equal-slope-across-factors* condition then yields the familiar one-parameter mean-variance benchmark. Alternatively, a demand-based extension of Ross (1976) Arbitrage Pricing Theory (APT) shows that when a handful of factors ($K \ll N$) explain most of the risk induced by trading demand, retaining only those K factors approximates the full N-factor model with negligible pricing error; introducing a single inverse-demand elasticity for the residual idiosyncratic shocks produces a practical $(K+1)$-parameter specification. Throughout, the analysis is agnostic about utility, payoff distributions, and market frictions; every result follows from no-arbitrage and the stated restrictions.

The result is an axiomatic anatomy of price impact—summarized in Figure 1—that cleanly separates what economic constraints must hold from what can be relaxed, and that offers a flexible middle ground between an empirically infeasible N^2 specification and a rigid one-parameter mean-variance benchmark.

Consider N non-redundant assets. Basic no-arbitrage reasoning imposes almost no struc-

ture on cross-asset impacts: demand for any one asset can influence the price of any other—its own included—creating an N^2 grid of potential interactions. Under a local linear approximation, any smooth impact curve can be summarized by its first-order slope at the observed shocks, so the entire grid is characterized by N^2 slope coefficients. The linear approximation serves only for simplicity; the dimensionality-reduction results can be generalized to nonlinear price-impact functions that satisfy the same no-arbitrage conditions.

At the opposite extreme, if the representative marginal investor clearing these demand shocks has mean-variance preferences, standard equilibrium results compress the N^2 price-impact coefficients into a single scalar—the investor's per-capita risk aversion—multiplied by the familiar return-covariance matrix. The resulting formula is widely used in the price-impact literature (see the survey by Rostek and Yoon, 2023). What bridges the gap between the N^2-parameter no-arbitrage representation and the one-parameter mean-variance benchmark? Equivalently, which economic conditions are both *necessary and sufficient* to shrink N^2 free coefficients to a single parameter?

The bridge is the Irrelevance of Uncorrelated Portfolios (IUP) axiom. IUP states that when two portfolios are uncorrelated in both fundamental risk and trading demand—whether directly or through any third portfolio—demand in one cannot move the price of the other. Empirical studies support the two channels that IUP rules out. First, arbitrage on shared fundamentals propagates shocks across assets that load on the same risk: for example, when noise traders buy airline stocks, marginal investors who sell those shares may buy oil-producer stocks to offset their fuel-price exposure, transmitting the shock across industries (Andrade, Chang, and Seasholes, 2008; Chaudhary, Fu, and Li, 2023). Second, common trading demand arises when investors transact in baskets of assets—such as through portfolio trades or exchange-traded funds (ETFs)—so demand for one stock echoes across the rest

of the basket (Hasbrouck and Seppi, 2001; Ben-David, Li, Rossi, and Song, 2022). IUP formalizes the converse: if both the arbitrage-driven and common-demand channels are absent, cross-impact must be zero.

The main theorem shows that the IUP axiom is both necessary and sufficient to collapse the full N^2-parameter representation to an N-parameter factor model. First, combine the N asset-level demands into N orthogonal factors, which are unique portfolios that are mutually uncorrelated in both fundamental risk and trading demand. Second, IUP implies that each factor's demand moves only its own price of risk; individual-asset prices adjust by no-arbitrage. Each factor $k = 1, \ldots, N$ has its own slope coefficient λ_k, which links demand to that factor's price of risk and captures the marginal risk-return trade-off faced by the investor who absorbs the demand. Unlike the standard inverse-demand elasticity—a direct quantity-to-price coefficient—λ_k is risk-based: it tracks how the extra risk created by the demand shock alters the compensation investors require.

The N-parameter factor model reduces to the one-parameter mean-variance benchmark when all factors share a common slope, $\lambda_1 = \lambda_2 = \cdots = \lambda_N$. Equal slopes follow from mean-variance preferences: a representative investor values an additional dollar of wealth identically, no matter which factor delivers it. The implication extends to CARA, CRRA, and recursive utility models: state variables can shift risk aversion over time, yet the local cross-section remains mean-variance, forcing a single λ.

Allowing the λ_k's to differ broadens the mean-variance benchmark from a single *marginal* risk-return trade-off to factor-specific ones—an extension that naturally parallels the long-standing evidence that *unconditional* risk-return trade-offs vary across factors (Fama and MacBeth, 1973). In the mean-variance benchmark a common λ equals the marginal investor's per-capita risk aversion; heterogeneous slopes would therefore require either factor-specific

risk aversion or segmentation that alters the effective investor base by factor. Standard preferences seldom generate such heterogeneity on their own, highlighting the value of an axiomatic, no-arbitrage approach and the role of the IUP axiom. Dynamic equilibrium work shows that distinct λ_k's can arise endogenously; for example, An and Zheng (2025) derive heterogeneous slopes in a CARA setting with predictable demand.

Instead of imposing equal slopes, one can reduce the N-factor system to a parsimonious K-factor model with $K \ll N$ by applying a demand-based APT. By construction, selecting the top K factors captures the maximum share of risk induced by trading demand. Hence, if the factor-level slopes λ_k are uniformly bounded, the remaining low-variance "tail" factors add only a negligible pricing error and can be dropped. The bounded-slope requirement acts as a "good-deal bound" in the sense of Cochrane and Saa-Requejo (2000): no factor can deliver an outsized marginal risk-return trade-off that arbitrageurs would quickly exploit.

The remaining $N-K$ portfolios represent the idiosyncratic demand specific to individual assets. The demand-based APT offers no structural guidance on how these residual shocks feed into prices. By construction these idiosyncratic demands are small and largely uncorrelated, making any cross-impact empirically hard to detect. A practical workaround treats each idiosyncratic shock as influencing only its own price and estimates a single inverse-demand elasticity, $\theta := (\Delta P_n/P_n)/(\Delta Q_n/Q_n)$, common across assets n. Empirical evidence that idiosyncratic demand exhibits an approximately constant price elasticity is provided by Davis (2025). The resulting specification therefore has K factor slopes, $\lambda_1, \ldots, \lambda_K$, and a single residual elasticity θ.

A natural follow-up question is whether one could skip the IUP step and jump straight from the full N^2 system to a K^2 alternative that keeps only the first K factors while still letting the retained factors cross-impact each other. Asset-pricing theory does not provide

a general way to do so without extra assumptions. When one computes the pricing error between the full N^2 specification and its K^2 truncation, a residual term remains: demand in the K retained factors still affects the prices of the $N-K$ discarded portfolios. Because those factor demands do not vanish, cross-impact can spill into the remaining assets. As a result, the residual cannot be bounded without explicitly assuming these cross-impacts are zero. In equity markets the boundary between systematic and idiosyncratic risk is often diffuse (see the "factor zoo" discussion in Cochrane, 2011), so assuming that cross-impact vanishes exactly at K must be supported by a defensible economic argument. The two-step path in Figure 1 sidesteps this issue: IUP first yields N orthogonal portfolios whose cross-impact is zero by construction, and the demand-based APT then selects the first K of these portfolios, mirroring the traditional APT choice of systematic factors without requiring cross-impact to vanish exactly at K.

Taken together, the results show that the IUP axiom untangles the N^2 system and links the mean-variance and factor representations. IUP identifies N unique portfolios that are mutually uncorrelated in fundamental risk and trading demand, thereby eliminating cross-impact among them. Although the construction is empirically motivated, the paper's second theorem gives a deeper theoretical rationale. Condensing an N^2 system to N parameters inevitably discards information. The theorem shows that, among all possible ways to select N portfolios and eliminate cross-impact, the IUP choice introduces the least distortion when moving from the full N^2 specification to the factor representation. Specifically, IUP is the only way that (i) leaves each portfolio's price sensitivity to its own demand unchanged, and (ii) leaves the marginal investor's expected utility unchanged.

Related literature. Asset-pricing research traditionally employs two complementary lenses: arbitrage arguments that rely solely on the absence of risk-free profit opportunities

(e.g., Markowitz, 1952; Black and Scholes, 1973; Ross, 1976), and equilibrium models that specify investors' preferences and endowments (e.g., Sharpe, 1964; Merton, 1973). For questions about how trading demands move prices, the equilibrium lens has so far dominated. Recent studies underscore the continued relevance of the arbitrage view. Kozak, Nagel, and Santosh (2018) show that competitive arbitrageurs can impose a factor structure on returns even in the presence of noise-trader demand, while Fuchs, Fukuda, and Neuhann (2025a) show how heterogeneous investor tastes can break no-arbitrage links. The present paper extends the arbitrage approach. Starting from no-arbitrage, it develops an axiomatic anatomy of price impact (Figure 1) that distinguishes economically necessary restrictions from those that are optional. The framework remains agnostic about utility, payoff distributions, and frictions, yet nests the usual equilibrium formulas as special cases, thereby helping to bridge reduced-form empirics and fully specified equilibria.

The analysis yields a parsimonious $K + 1$ specification, intermediate between the full N^2 no-arbitrage representation and the one-parameter mean-variance benchmark. Unlike demand-system models (e.g., Koijen and Yogo, 2019), mean-variance frameworks (e.g., Vayanos and Vila, 2021), or hybrid variants (e.g., Jansen, Li, and Schmid, 2025), this approach lets the data identify the dominant K factors, applies the IUP axiom to orthogonalize them, and assigns each a distinct marginal price-of-risk slope λ_k. All remaining, largely uncorrelated shocks are pooled into a single inverse-demand elasticity θ. Haddad, He, Huebner, Kondor, and Loualiche (2025) work with a researcher-selected set of K characteristic portfolios and estimate a $K^2 + 1$ elasticity system in which each retained characteristic can influence the others. Both methods have empirical appeal. The specification developed here emphasizes parsimony and, through IUP orthogonalization, gives each coefficient—especially the factor-level slopes λ_k—a clear economic interpretation. Empirical implementations in equities and

foreign exchange are provided by An, Su, and Wang (2025) and An and Huber (2025), respectively.

Road map. Section 2 introduces the market environment and notation. Section 3 states the axioms and shows how the IUP axiom collapses the N^2 no-arbitrage representation to an N-parameter factor model. Section 4 interprets the factor-specific slopes λ_k and shows how low-dimensional demand leads to a $K+1$ specification. Section 5 proves that IUP provides the minimally distortive transformation and extends the framework to nonlinear impact curves. Section 6 concludes. The appendices contain proofs and additional results.

2 Model Setup

Time is discrete with two periods, $t = 0$ and $t = 1$. For each period $t \in \{0, 1\}$, let $(\Omega_t, \mathcal{F}_t, \mathbb{P}_t)$ denote the underlying probability space. Define the product space $(\Omega, \mathcal{F}, \mathbb{P}) := (\Omega_0 \times \Omega_1, \mathcal{F}_0 \times \mathcal{F}_1, \mathbb{P}_0 \times \mathbb{P}_1)$, which supports all random variables introduced below. Arguments (ω_0, ω_1) are displayed only when needed. Boldface letters denote vectors and matrices; $\mathbf{A}^\top$ is the transpose of $\mathbf{A}$. The remainder of this section introduces asset payoffs, trading-demand shocks, and the stochastic discount factor (SDF) used throughout the analysis.

Consider N assets indexed by $n = 1, \ldots, N$. Asset n has an outstanding market value of Q_n dollars. At $t = 0$ a group of noise traders places an order demanding $\Delta Q_n(\omega_0)$ dollars of asset n, which a representative marginal investor absorbs. Collect these demands as $\Delta\mathbf{Q} = (\Delta Q_1, \ldots, \Delta Q_N)^\top$. All amounts are expressed in dollars relative to the no-shock price $P_n(\mathbf{0})$ introduced below.[1] Assume the demand shocks are mean-zero, $\mathbb{E}[\Delta\mathbf{Q}] = \mathbf{0}$, so the analysis focuses on their variation. If a non-zero mean is empirically relevant, all results

[1] Dollar units make shocks directly comparable even when assets differ in share price or share count. For example, purchasing two A-shares at \$50 and one B-share at \$100 both register as a \$100 demand. Dollar scaling is therefore invariant to mechanical actions such as stock splits, which alter share count but not trade size, and is standard in empirical work.

continue to hold after switching from the covariance matrix $\text{var}(\Delta\mathbf{Q})$ to the second moment $\mathbb{E}[\Delta\mathbf{Q}\Delta\mathbf{Q}^\top]$, analogous to using raw rather than demeaned data in principal-component analysis.[2]

At $t = 1$ each asset n delivers a payoff $X_n(\omega_1)$, collected as $\mathbf{X} = (X_1, \ldots, X_N)^\top$. The trading demands are assumed to be uninformed: $\Delta\mathbf{Q}$ is independent of $\mathbf{X}$. The gross risk-free rate, R_F, is constant and unaffected by trading demand. Assume $\text{var}(\mathbf{X})$ and $\text{var}(\Delta\mathbf{Q})$ are full rank; if not, one could first extract linearly independent portfolios and rotate $\mathbf{X}$ and $\Delta\mathbf{Q}$ accordingly. No parametric distributional assumptions are imposed on $\mathbf{X}$ or $\Delta\mathbf{Q}$.

Let $\mathbf{P}(\Delta\mathbf{Q}) = (P_1(\Delta\mathbf{Q}), \ldots, P_N(\Delta\mathbf{Q}))^\top$ denote time-0 asset prices. Prices adjust when the representative marginal investor absorbs the trading demand $\Delta\mathbf{Q}$. Except for the demand shock $\Delta\mathbf{Q}$, the framework follows the textbook arbitrage setup in Cochrane (2009).

The first assumption is the law of one price (LOOP). Conditional on any $\Delta\mathbf{Q}$, two assets that deliver identical payoffs at $t = 1$ must command the same price at $t = 0$. LOOP ensures that no risk-free arbitrage remains after prices adjust to the demand shock. This assumption contrasts with the "noise-trader risk" model of De Long, Shleifer, Summers, and Waldmann (1990), where noise trading is an additional risk factor and identical payoffs need not imply identical prices. Under LOOP, for any trading demand $\Delta\mathbf{Q}$ there exists a stochastic discount factor (SDF) $M(\Delta\mathbf{Q})$ satisfying[3]

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

The expectation on the right-hand side is taken over the time-1 uncertainties in $M(\Delta\mathbf{Q})$

[2]Some studies (Hartzmark and Solomon, 2025; Vayanos, 2021) show that the expected component of demand shocks can also move prices.

[3]Because redundant assets are removed by the full-rank assumption on $\text{var}(\mathbf{X})$, deriving equation (1) does not require LOOP. However, LOOP is still assumed because equation (1) is later used to price portfolios, which are redundant given the underlying assets.

and $\mathbf{X}$, conditional on the realized demand $\Delta\mathbf{Q}$.[4] Whether $\Delta\mathbf{Q}$, $\mathbf{P}(\Delta\mathbf{Q})$, $M(\Delta\mathbf{Q})$, and $\mathbf{X}$ are exogenous or endogenous is immaterial; equation (1) holds in any equilibrium that satisfies LOOP. The pricing relation therefore characterizes asset values across a wide range of equilibria. When $\Delta\mathbf{Q} = \mathbf{0}$,

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

This no-shock price vector serves as the baseline for the impact analysis that follows.

For any trading demand $\Delta\mathbf{Q}$, define the percentage price impact

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

The fundamental return—the return that would obtain absent demand shocks—is

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

Because the denominators are the no-shock prices $\mathbf{P}(\mathbf{0})$, randomness in $\mathbf{R}_0$ stems solely from payoff uncertainty in $\mathbf{X}$.

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

$$\Delta\mathbf{p}(\Delta\mathbf{Q}) = \mathbb{E}[\Delta M(\Delta\mathbf{Q})\mathbf{R}_0], \tag{5}$$

where

$$\Delta M(\Delta\mathbf{Q}) := M(\Delta\mathbf{Q}) - M(\mathbf{0}) \tag{6}$$

is the change in the SDF induced by the trading demand $\Delta\mathbf{Q}$. Equation (5) shows that

[4] To be precise, the SDF $M(\Delta\mathbf{Q})$ is a random variable $M(\Delta\mathbf{Q}(\omega_0), \omega_1)$, which depends on the realization of trading demand $\Delta\mathbf{Q}(\omega_0)$ at time 0 and also incorporates the payoff randomness of ω_1 at time 1.

characterizing the multi-asset price-quantity relation $\Delta\mathbf{p}(\Delta\mathbf{Q})$ requires only the marginal response $\Delta M(\Delta\mathbf{Q})$, not the full functional form of $M(\cdot)$. Economically, $\Delta M(\Delta\mathbf{Q})$ captures how trading demand alters the marginal investor's pricing of risk.

Because the risk-free rate R_F is constant and unaffected by trading demand, the SDF must satisfy $\mathbb{E}[M(\Delta\mathbf{Q})R_F] = 1$ for every $\Delta\mathbf{Q}$. Hence the change in the SDF has zero mean:

$$\mathbb{E}[\Delta M(\Delta\mathbf{Q})] = 0. \tag{7}$$

In summary, price impacts $\Delta\mathbf{p}(\Delta\mathbf{Q})$ are linked to the SDF change $\Delta M(\Delta\mathbf{Q})$ through equation (5). The next section introduces a set of axioms that pin down the structure of $\Delta M(\Delta\mathbf{Q})$ and, by extension, the form of $\Delta\mathbf{p}(\Delta\mathbf{Q})$.

3 Axiomatic Foundations and Factor Representation

This section organizes the theory in four steps. Section 3.1 introduces three baseline axioms and derives an $N \times N$ price-impact representation with N^2 free parameters. Section 3.2 shows that, under standard CARA utility and jointly normal payoffs, the general N^2 no-arbitrage model reduces to the familiar one-parameter mean-variance formula. To bridge the gap between an N^2 specification and a one-parameter form, Section 3.3 adds one more axiom—the *Irrelevance of Uncorrelated Portfolios* (IUP). IUP reduces the representation to an N-parameter factor structure in which each factor's demand shifts only its own price of risk. Finally, Section 3.4 shows that imposing an *equal-slope-across-factors* condition on the factor model recovers the one-parameter mean-variance benchmark.

3.1 Baseline Axioms and the N^2 Representation

ASSUMPTION 1. ***Linearity of Price Impact***

For any scalars a_1, a_2 and vectors $\Delta\mathbf{Q}_1, \Delta\mathbf{Q}_2 \in \mathbb{R}^N$,

$$\Delta\mathbf{p}(a_1\Delta\mathbf{Q}_1 + a_2\Delta\mathbf{Q}_2) = a_1\Delta\mathbf{p}(\Delta\mathbf{Q}_1) + a_2\Delta\mathbf{p}(\Delta\mathbf{Q}_2). \tag{8}$$

Linearity is intended as a local approximation: over a small neighborhood of the data, any smooth impact curve can be summarized by its slope. For this reason empirical work typically estimates price impact with linear regressions around the observed shocks. Section 5.2 shows that the main results extend when the linearity assumption is relaxed.

ASSUMPTION 2. ***Positive Compensation for Risk***

For any non-zero trading demand $\Delta\mathbf{Q} \neq \mathbf{0}$, $\Delta\mathbf{Q}^\top\Delta\mathbf{p}(\Delta\mathbf{Q}) > 0$.

With a single asset, the condition $\Delta Q \cdot \Delta p(\Delta Q) > 0$ says that a positive demand raises the price and a negative demand lowers it, so the marginal investor earns positive compensation for supplying or absorbing risk. With multiple assets the dot product $\Delta\mathbf{Q}^\top\Delta\mathbf{p}(\Delta\mathbf{Q}) > 0$ requires the price-impact vector to move in the same overall direction as the demand vector, extending the same intuition to a portfolio of trades.

ASSUMPTION 3. ***Projection of the SDF***

For any trading demand $\Delta\mathbf{Q}$, the change in the SDF $\Delta M(\Delta\mathbf{Q})$ lies in the space spanned by demeaned fundamental returns $\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]$.

Hansen and Richard (1987) show that the SDF is unique within the return space; requiring the same projection for $\Delta M(\Delta\mathbf{Q})$ therefore preserves identification. Because $\mathbb{E}[\Delta M(\Delta\mathbf{Q})] = 0$ in (7), the projection is onto the mean-zero space generated by $\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]$.

Taken together, the three axioms characterize both the SDF change $\Delta M(\Delta \mathbf{Q})$ and the resulting price impact $\Delta \mathbf{p}(\Delta \mathbf{Q})$. Appendix B.1 provides a proof.

PROPOSITION 1. *Assumptions 1, 2, and 3 hold if and only if*

$$\Delta M(\Delta \mathbf{Q}) = (\mathbf{Y}\Delta \mathbf{Q})^{\top}(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]), \tag{9}$$

where $\mathbf{Y}$ *is an* $N \times N$ *parameter matrix such that* $\mathrm{var}(\mathbf{R}_0)\mathbf{Y}$ *is positive definite. The corresponding price-impact representation is*

$$\Delta \mathbf{p}(\Delta \mathbf{Q}) = \mathbb{E}[\Delta M(\Delta \mathbf{Q})\mathbf{R}_0] = \mathrm{var}(\mathbf{R}_0)\mathbf{Y}\Delta \mathbf{Q}. \tag{10}$$

Expression (10) links fundamental risk, $\mathrm{var}(\mathbf{R}_0)$, to trading demand, $\Delta \mathbf{Q}$, through an $N \times N$ matrix $\mathbf{Y}$. With N^2 free coefficients in $\mathbf{Y}$, the baseline axioms leave a system that is impractical to estimate directly. Two complementary restrictions can shrink this dimensionality: (i) an equilibrium assumption that collapses all N^2 parameters to a single mean-variance coefficient (next subsection), and (ii) an arbitrage-based axiom that reduces the system to N parameters (Section 3.3).

3.2 Mean-Variance Price-Impact Model

Proposition 1 shows that the three baseline axioms alone generate an N^2-parameter price-impact representation. If, in addition, one imposes the standard representative-agent equilibrium assumptions—CARA preferences and jointly normal payoffs—the same representation collapses to the familiar one-parameter mean-variance formula: the matrix $\mathbf{Y}$ becomes a scalar multiple of the identity matrix. Thus the mean-variance model is a special case contained within the broader N^2 framework, illustrating that an extra axiom is required to

obtain the intermediate N-parameter factor specification introduced later.

Consider a representative marginal investor of mass μ with CARA utility $u(c) = -e^{-\gamma c}$. Let the payoff vector $\mathbf{X}$ be jointly normal. The normality assumption is not essential; a log-linear approximation around the mean yields the same closed-form expressions.

In equilibrium, prices $\mathbf{P}(\Delta\mathbf{Q})$ adjust until the representative marginal investor is willing to absorb the shock $-\Delta\mathbf{Q}/\mu$ dollars—equivalently $-\Delta\mathbf{Q}/(\mathbf{P}(\mathbf{0})\mu)$ shares (element-wise division)—so that markets clear:

$$-\frac{\Delta\mathbf{Q}}{\mathbf{P}(\mathbf{0})\mu} = \arg\max_{\mathbf{y}} \mathbb{E}[-\exp(-\gamma\left(\mathbf{y}^\top(\mathbf{X} - \mathbf{P}(\Delta\mathbf{Q})R_F) + \mathbf{S}^\top\mathbf{X}\right))]. \tag{11}$$

Here the investor selects the share vector $\mathbf{y}$ at $t = 0$, pays $\mathbf{P}(\Delta\mathbf{Q})$ per share, finances that cost at the gross risk-free rate R_F, and receives $\mathbf{X}$ at $t = 1$. The vector $\mathbf{S}$ records the investor's pre-shock holdings. The first-order condition of (11) implies

$$\mathbf{P}(\Delta\mathbf{Q}) = \frac{\mathbb{E}[\mathbf{X}]}{R_F} - \frac{\gamma}{R_F}\mathrm{var}(\mathbf{X})\left(\mathbf{S} - \frac{\Delta\mathbf{Q}}{\mathbf{P}(\mathbf{0})\mu}\right). \tag{12}$$

Subtracting the $\Delta\mathbf{Q} = \mathbf{0}$ case from the general case and dividing element-wise by $\mathbf{P}(\mathbf{0})$ gives, via definitions (3) and (4),

$$\Delta\mathbf{p}(\Delta\mathbf{Q}) = \frac{\gamma}{\mu R_F}\mathrm{var}(\mathbf{R}_0)\Delta\mathbf{Q}. \tag{13}$$

This is the standard mean-variance price-impact formula (see Rostek and Yoon, 2023).

Equation (13) is a special case of Proposition 1: it corresponds to $\mathbf{Y} = (\gamma/\mu R_F)\,\mathbf{I}_N$, where $\mathbf{I}_N$ is the $N \times N$ identity matrix. This equilibrium choice collapses the N^2 free parameters to a single scalar, γ/μ, the marginal investor's per-capita risk aversion. Section 3.3

next introduces the Irrelevance of Uncorrelated Portfolios (IUP) axiom, reducing the N^2 representation to an N-parameter factor form. Section 3.4 then shows when that factor form further contracts to the one-parameter mean-variance formula (13). In short, the mean-variance model provides the extreme $N^2 \to 1$ compression, whereas the IUP axiom delivers the intermediate, empirically flexible $N^2 \to N$ step.

3.3 IUP and the Factor Representation

With the $N^2 \to 1$ mean-variance benchmark in place, this subsection shows how the IUP axiom reduces the baseline N^2-parameter representation to an N-parameter factor model. The argument proceeds in four parts:

(i) Section 3.3.1 states and motivates the IUP axiom (Assumption 4).

(ii) Section 3.3.2 introduces a regularity condition (Assumption 5) that simplifies the main text; cases without this condition are treated in Online Appendix A.

(iii) Section 3.3.3 constructs N canonical factors—portfolios orthogonal in both fundamental returns and trading demand (Proposition 2).

(iv) Section 3.3.4 combines the ingredients above to derive the factor representation, in which each factor's demand affects only its own price of risk (Theorem 1).

3.3.1 IUP Axiom

Empirical research supports two channels through which demand for one asset (e.g., Google) can influence the price of another (e.g., Meta). First, *correlated fundamental risk*. When noise traders buy Google, marginal investors sell Google to clear the market. To offset the

resulting technology-sector exposure, they may simultaneously buy Meta. This arbitrage-based rebalancing transmits the original shock across assets and is documented by Andrade, Chang, and Seasholes (2008) and Chaudhary, Fu, and Li (2023). Second, *correlated trading demand*. Investors frequently trade Google and Meta together—through baskets or ETFs—so demand for one stock echoes across the rest of the basket, a pattern documented by Hasbrouck and Seppi (2001) and Ben-David, Li, Rossi, and Song (2022).[5]

The IUP axiom formalizes the converse. When two portfolios are uncorrelated in both fundamental risk and trading demand—directly or through any third portfolio—demand in one cannot influence the price of the other. The following statement makes the axiom precise.

ASSUMPTION 4. ***Irrelevance of Uncorrelated Portfolios (IUP)***

Let s be the trading-demand shock to a portfolio with weight vector

$$\mathbf{d} = \left(\mathrm{cov}(\Delta Q_1, s), \mathrm{cov}(\Delta Q_2, s), \ldots, \mathrm{cov}(\Delta Q_N, s)\right)^\top, \tag{14}$$

and let $\mathbf{a} \in \mathbb{R}^N$ be any portfolio weight vector. If $\mathbf{a}^\top \mathbf{C}\mathbf{d} = 0$ for any matrix $\mathbf{C}$ satisfying

$$\mathrm{var}(\mathbf{R}_0)\mathrm{var}(\Delta\mathbf{Q})\mathbf{C} = \mathbf{C}\mathrm{var}(\Delta\mathbf{Q})\mathrm{var}(\mathbf{R}_0), \tag{15}$$

then $\mathrm{cov}(\mathbf{a}^\top \Delta\mathbf{p}(\Delta\mathbf{Q}), s) = 0$.

In words, demand for portfolio $\mathbf{d}$ leaves the price of portfolio $\mathbf{a}$ unaffected whenever the two portfolios are uncorrelated along both risk and demand dimensions. Here s is the trading demand shock for portfolio $\mathbf{d}$, and $\mathbf{a}^\top \Delta\mathbf{p}(\Delta\mathbf{Q})$ is the corresponding price change for portfolio $\mathbf{a}$. The zero covariance condition $\mathrm{cov}(\mathbf{a}^\top \Delta\mathbf{p}(\Delta\mathbf{Q}), s) = 0$ therefore rules out cross-impact

[5]Such trading comovement can arise from shared behavioral biases or from investors facing common liquidity shocks; see Caballe and Krishnan (1994) and Kumar and Seppi (1994).

between uncorrelated portfolios.

The term "uncorrelated" must rule out not only direct links between two portfolios, $\mathbf{a}$ and $\mathbf{d}$, but also any indirect link that could arise through a third portfolio. In a multi-asset setting this is subtle: two portfolios can have zero pair-wise covariances in both fundamental risk and trading demand, yet still be connected once a third portfolio is taken into account. A precise statement is therefore needed.

The formal test uses the fact that two diagonalizable matrices commute—that is, $\mathbf{AB} = \mathbf{BA}$—if and only if they can be made diagonal by the same rotation. Here the relevant matrices are the fundamental-return covariance $\text{var}(\mathbf{R}_0)$ and the demand covariance $\text{var}(\Delta\mathbf{Q})$. Any matrix $\mathbf{C}$ satisfying $\text{var}(\mathbf{R}_0)\text{var}(\Delta\mathbf{Q})\mathbf{C} = \mathbf{C}\text{var}(\Delta\mathbf{Q})\text{var}(\mathbf{R}_0)$ in (15) captures a possible indirect connection between $\mathbf{a}$ and $\mathbf{d}$ through some third portfolio. Requiring $\mathbf{a}^\top\mathbf{C}\mathbf{d} = 0$ for every such $\mathbf{C}$ guarantees that $\mathbf{a}$ and $\mathbf{d}$ remain orthogonal—simultaneously—along both the fundamental risk and demand dimensions, even after all potential intermediate portfolios are accounted for. In this sense the two portfolios lie in distinct risk-and-demand directions, and the IUP axiom declares that demand in one cannot influence the price of the other.

A brief remark explains why s is defined as the trading-demand shock for portfolio $\mathbf{d}$ in (B.15). Suppose the shock s rises by one dollar. The associated change in the demand for asset n is the regression coefficient of ΔQ_n on s, namely $\text{cov}(\Delta Q_n, s)/\,\text{var}(s)$. At the same time, the weight vector $\mathbf{d} = (d_1, \ldots, d_N)^\top$ describes how a one-dollar demand for portfolio $\mathbf{d}$ is distributed across individual assets: asset n receives d_n dollars. Choosing $d_n = \text{cov}(\Delta Q_n, s)/\,\text{var}(s)$ therefore makes the portfolio's dollar demand exactly replicate the shock s. Equation (B.15) multiplies these weights by the scalar $\text{var}(s)$ purely for notational convenience; the scale factor does not affect any covariance conditions that follow.

3.3.2 Technical Regularity

With IUP in place, the next task is to characterize the SDF change $\Delta M(\Delta \mathbf{Q})$ and the price impact $\Delta \mathbf{p}(\Delta \mathbf{Q})$. The following assumption simplifies the exposition.

ASSUMPTION 5. ***Distinct Eigenvalues***

Let the Cholesky decomposition of the fundamental risk matrix be $\text{var}(\mathbf{R}_0) = \mathbf{U}^\top \mathbf{U}$ *with* $\mathbf{U}$ *upper triangular. Assume the matrix* $\text{var}(\mathbf{U}\Delta \mathbf{Q})$ *has distinct eigenvalues.*

In linear-algebra terms, distinct eigenvalues represent the generic case; repeated eigenvalues are a knife-edge situation handled in Online Appendix A. When eigenvalues repeat, IUP imposes no restriction on cross-impact within the associated eigenspace, because commonality in risk and demand fails to single out a unique set of N factors. Economically, this means the data do not fully identify the price-impact structure, and the resulting model is correspondingly less restrictive.

3.3.3 Constructing Canonical Factors

Assumption 5 sets the stage for an explicit factor structure. The next lemma shows that—regardless of the correlation patterns in $\text{var}(\Delta \mathbf{Q})$ and $\text{var}(\mathbf{R}_0)$—one can extract a unique set of N portfolios whose fundamental returns and trading demands are mutually uncorrelated. Under IUP these portfolios have zero cross-impact and therefore serve as the *canonical factors* governing both the SDF change $\Delta M(\Delta \mathbf{Q})$ and the price-impact map $\Delta \mathbf{p}(\Delta \mathbf{Q})$. Proposition 2, proved in Appendix B.2, formalizes this construction.

PROPOSITION 2. *There exist* N *canonical factors—unique up to sign—whose portfolio weight vectors are* $\mathbf{w}_k = (w_{1,k}, w_{2,k}, \ldots, w_{N,k})^\top$ *for* $k = 1, 2, \ldots, N$. *They satisfy:*

i. Factor-demand decomposition:

$$\Delta\mathbf{Q} = \sum_{k=1}^{N} \mathbf{w}_k \Delta Q_k^F, \quad \textit{equivalently,} \quad \Delta Q_n = \sum_{k=1}^{N} w_{n,k} \Delta Q_k^F, \tag{16}$$

where ΔQ_k^F is the trading demand for factor k.

ii. Uncorrelated factor returns:

$$\operatorname{cov}(\mathbf{w}_k^\top \mathbf{R}_0, \mathbf{w}_j^\top \mathbf{R}_0) = 0 \textit{ for all } k \neq j, \textit{ and } \operatorname{var}(\mathbf{w}_k^\top \mathbf{R}_0) = 1 \textit{ for all } k. \tag{17}$$

iii. Uncorrelated factor demand:

$$\operatorname{cov}(\Delta Q_k^F, \Delta Q_j^F) = 0 \textit{ for all } k \neq j, \textit{ and } \operatorname{var}(\Delta Q_1^F) > \operatorname{var}(\Delta Q_2^F) > \cdots > \operatorname{var}(\Delta Q_N^F). \tag{18}$$

Equation (16) is bookkeeping: a \$1 demand for factor k allocates $w_{n,k}$ dollars to asset n; summing across factors reproduces each asset's total demand. Conditions (17) and (18) then pin down the factor directions. Condition (17) makes the N factors' fundamental returns mutually uncorrelated and scaled to unit variance. Condition (18) makes the corresponding factor demands mutually uncorrelated and orders them by descending demand variance, $\operatorname{var}(\Delta Q_1^F) > \operatorname{var}(\Delta Q_2^F) > \cdots > \operatorname{var}(\Delta Q_N^F)$. These two orthogonality requirements leave only a sign ambiguity, so the weight vectors $\{\mathbf{w}_k\}_{k=1}^N$ are unique up to sign flips.

The canonical factors resemble principal-component analysis (PCA) but differ in a crucial respect. A standard PCA is applied to *either* returns *or* trading quantities, never to both together. Run on returns, PCA enforces condition (17), mutual orthogonality of factor

returns, and then fixes a rotation to identify the factors uniquely.[6] Run on quantities, PCA imposes only condition (18), orthogonality of factor demands. By contrast, the canonical factors satisfy *both* conditions simultaneously: their returns are uncorrelated *and* their demands are uncorrelated. One can view the construction as a single, modified PCA applied jointly to returns and demands, rather than two separate decompositions.[7] This joint orthogonality is essential for eliminating cross-factor impacts under IUP. Using foreign-exchange data, An and Huber (2025) show that the joint procedure selects factors with sharper economic interpretation and yields materially different results from either one-sided PCA, underscoring its empirical value.

3.3.4 Factor Representation Theorem

With the canonical factors in place, the entire price-impact system can now be written in factor form. The following theorem—central to the paper—shows that, under IUP, both the SDF change and the price-impact map admit a parsimonious representation in the canonical factors identified in Proposition 2. A proof appears in Appendix B.3.

THEOREM 1. *Assumptions 1, 2, 3, and 4 hold if and only if the change in the SDF can be written*

$$\Delta M(\Delta \mathbf{Q}) = \sum_{k=1}^{N} \lambda_k \Delta Q_k^F \mathbf{w}_k^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) \tag{19}$$

for positive parameters $\lambda_1, \ldots, \lambda_N$. The corresponding price-impact relation is

$$\Delta \mathbf{p}(\Delta \mathbf{Q}) = \mathbb{E}[\Delta M(\Delta \mathbf{Q}) \mathbf{R}_0] = \sum_{k=1}^{N} \lambda_k \Delta Q_k^F \operatorname{cov}(\mathbf{R}_0, \mathbf{w}_k^\top \mathbf{R}_0). \tag{20}$$

[6] The usual convention is to impose orthogonal loadings; see Jolliffe (1986).

[7] Gârleanu and Pedersen (2022) analyze the factor structure of returns and of trading quantities but treat the two decompositions separately.

To interpret (19), recall from Hansen and Richard (1987) that projecting the baseline SDF onto the return space (Assumption 3) gives:

$$M(\mathbf{0}) = \frac{1}{R_F} + \sum_{k=1}^{N} \psi_k \mathbf{w}_k^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]), \tag{21}$$

where $1/R_F$ is the mean of the SDF, $\mathbf{w}_k^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])$ is the demeaned return on factor k, and ψ_k is that factor's price of risk. Comparing $\Delta M(\Delta\mathbf{Q}) = M(\Delta\mathbf{Q}) - M(\mathbf{0})$ in (19) with $M(\mathbf{0})$ in (21) shows that by linearity (Assumption 1), a demand shock ΔQ_k^F changes the price of risk with slope λ_k. Positive-compensation (Assumption 2) guarantees $\lambda_k > 0$ for every factor. Finally, the IUP axiom (Assumption 4) rules out cross-effects: demand for factor k alters only its own price of risk, so mixed terms such as $\Delta Q_k^F \mathbf{w}_j^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])$ with $k \neq j$ cannot appear.

The implied price impacts in (20) inherit the orthogonality of the factors. Because the factors are constructed to have uncorrelated fundamental returns (condition (17)), the price impact of factor k is given by

$$\mathbf{w}_k^\top \Delta\mathbf{p}(\Delta\mathbf{Q}) = \lambda_k \Delta Q_k^F \operatorname{var}(\mathbf{w}_k^\top \mathbf{R}_0). \tag{22}$$

Thus the price change for factor k depends only on its own trading demand ΔQ_k^F and its own fundamental variance; other factors are absent. This no-cross-impact property—imposed by IUP—cuts the parameter count from N^2 (the matrix $\mathbf{Y}$ in (10)) to N slopes $\lambda_1, \ldots, \lambda_N$. Section 4.1 interprets these slopes economically.

Finally, equation (22) translates to individual assets via the law of one price: each asset's impact is the sum of its covariances $\operatorname{cov}(\mathbf{R}_0, \mathbf{w}_k^\top \mathbf{R}_0)$ with the factor returns, scaled by the corresponding λ_k and demand shock ΔQ_k^F.

Empirical work often relies on pre-defined portfolios—such as the Fama–French factors—that are not necessarily orthogonal in either returns or demands. Using such factors re-introduces an $N \times N$ cross-impact matrix because the factors generally influence one another. Even so, the IUP axiom keeps the system parsimonious: the cross-impact matrix for any such factor basis is still governed by only N free parameters rather than N^2. Appendix A spells out the resulting structural restrictions.

3.4 Mean-Variance as a Special Case: Equal Slopes Across Factors

The single-parameter mean-variance model in Section 3.2 nests inside the N-parameter factor model of Section 3.3 once an *equal-slope-across-factors* restriction is imposed, $\lambda_1 = \cdots = \lambda_N$.

Insert the factor-demand decomposition (16) into the mean-variance impact formula (13):

$$\Delta \mathbf{p}(\Delta \mathbf{Q}) = \sum_{k=1}^{N} \frac{\gamma}{\mu R_F} \Delta Q_k^F \operatorname{cov}(\mathbf{R}_0, \mathbf{w}_k^\top \mathbf{R}_0). \tag{23}$$

Expression (23) is identical to the factor representation (20) with the common slope $\lambda_k = \gamma/(\mu R_F)$ for every $k = 1, 2, \ldots, N$, where γ is the representative investor's risk-aversion coefficient, μ the investor mass, and R_F the gross risk-free rate.

Hence, in equilibrium the arbitrage slope λ_k equals the representative investor's per-capita risk aversion, γ/μ. The equal-slope condition embodies mean-variance utility: an additional dollar of wealth is valued the same way regardless of which factor delivers it. The argument extends to CRRA and recursive preferences as well—state variables can shift aggregate risk aversion over time, but the instantaneous cross-section remains mean-variance, forcing a common λ_k across all factors.

Allowing the slopes λ_k to differ generalizes the mean-variance benchmark from one

marginal risk-return trade-off to factor-specific trade-offs. This extension mirrors long-standing evidence that *unconditional* risk-return relationships vary across factors (Fama and MacBeth, 1973). Once the baseline SDF in (21) assigns each factor its own unconditional price of risk ψ_k, it is natural to let the demand sensitivity of that price—λ_k—vary as well. In the mean-variance benchmark, all slopes equal $\lambda = \gamma/(\mu R_F)$; heterogeneity would therefore require factor-specific risk aversion γ or segmentation that alters the effective investor mass μ by factor. Because standard preferences seldom generate such heterogeneity on their own, the arbitrage route—anchored by the IUP axiom—remains essential for the N-slope representation.

Although the mean-variance lens suggests reading λ_k as a factor-specific risk-aversion coefficient, the parameter is determined solely by the paper's axioms, which impose no assumptions on utility, payoff distributions, or trading frictions. It is therefore best viewed simply as the slope linking demand for factor k to that factor's price of risk. In practice this slope may reflect classical risk aversion—as in the mean-variance benchmark—or other market features, such as investment mandates that limit how marginal investors adjust exposure to particular factors (Gabaix and Koijen, 2022).

To summarize, the dimensionality ladder now reads $N^2 \to N \to 1$. Progress along that ladder turns on two choices: which factors to retain and whether the slopes λ_k are allowed to differ across them. The next section investigates the economic forces that govern these slopes and shows how they justify trimming the N-factor model to a parsimonious K-factor specification.

4 Economic Meaning and Dimensionality Reduction

Section 3 reduced the price-quantity system to the N slopes λ_k. Section 4.1 clarifies the economic content of these slopes, distinguishing them from a factor's unconditional price of risk and from the usual inverse-demand elasticity. Section 4.2 then asks how far dimensionality can fall: under what conditions can a demand-based APT justify keeping only the $K \ll N$ most important factors, yielding a parsimonious $(K+1)$-parameter specification that combines those K slopes with a single residual elasticity. Finally, Section 4.3 explains why an apparently natural K^2 truncation—allowing all pairwise interactions among the first K factors—fails without the IUP axiom, underscoring that axiom's role. Together, these results extend the classic APT to settings with correlated demand and delineate which dimension-reduction shortcuts are theoretically sound.

4.1 Price Sensitivity to Risk: Interpreting the λ_k Slopes

Equation (22) expresses each slope as

$$\lambda_k = \frac{\mathbf{w}_k^\top \Delta\mathbf{p}(\Delta\mathbf{Q})}{\Delta Q_k^F \mathrm{var}(\mathbf{w}_k^\top \mathbf{R}_0)} = \frac{\text{change in factor return induced by demand}}{\text{change in quantity of factor risk induced by demand}}. \tag{24}$$

The numerator $\mathbf{w}_k^\top \Delta\mathbf{p}(\Delta\mathbf{Q})$—the change in its return induced by the demand shock. The denominator $\Delta Q_k^F \mathrm{var}(\mathbf{w}_k^\top \mathbf{R}_0)$ is the incremental amount of factor risk induced by the same shock. Hence λ_k is the factor's *price sensitivity to risk*: the compensation the marginal investor requires per additional unit of factor risk induced by trading demand.

The price sensitivity to risk λ_k connects two familiar concepts. In standard asset-pricing theory a factor's *price of risk*—its risk premium divided by its quantity of risk—summarizes the unconditional risk-return trade-off facing the representative investor. Here, because the

demand shock ΔQ_k^F carries no information about time-1 payoffs, any price impact at $t = 0$ must unwind by $t = 1$; the numerator in (24) is therefore exactly the demand-induced change in factor k's risk premium over that interval. Accordingly, λ_k is a *marginal* price of risk—conditional on the trading demand—set by the investor who absorbs the incremental trade and prices the incremental risk.

Second, compare λ_k with the inverse-demand elasticity used in industrial-organization studies,

$$\varepsilon = (\Delta P/P)/(\Delta Q/Q), \tag{25}$$

which scales the price change by the percentage change in the *quantity of securities*. In contrast, λ_k scales the price change by the change in the *quantity of risk*. The distinction is economic: suppose the marginal investor absorbs equal dollar buy orders in two assets. If the assets' returns are perfectly negatively correlated, the combined position is riskless and requires no compensation, even though the share quantities have changed—so ε says nothing about the premium. If the correlation is imperfect, the position adds non-diversifiable risk and the investor demands compensation, captured by λ_k rather than by ε. Thus the relevant "elasticity" for asset pricing should be defined with respect to risk introduced by trading demand, not the sheer number of shares traded.

The inverse-demand elasticity ε uses a *percentage* change in the quantity of securities in its denominator, making the measure scale-free so that a \$1 demand for a small-cap stock is comparable to a \$1 demand for a large-cap stock (Davis, 2025). In contrast, the denominator of λ_k is the *absolute* change in the quantity of factor risk, because the aim is to compare slopes across factors—portfolios that can be levered arbitrarily. This choice renders λ_k invariant to portfolio leverage. Doubling every position in factor k (that is, scaling each weight $w_{n,k}$ by 2) doubles the factor's price impact $\mathbf{w}_k^\top \Delta\mathbf{p}(\Delta\mathbf{Q})$, halves the associated

factor demand ΔQ_k^F because of decomposition (16), and quadruples the return variance $\text{var}(\mathbf{w}_k^\top \mathbf{R}_0)$; the three changes offset exactly, leaving λ_k unchanged. The slope thus provides an apples-to-apples comparison across factors regardless of their leverage.

4.2 A Demand-Based APT

Equation (20) includes all N factors, yet in practice only a handful account for most of the risk introduced by trading demand. This subsection extends Ross's APT to show when the N-slope system can be trimmed to $K \ll N$ factors. Under the stated conditions the contributions of the remaining $N - K$ "tail" factors become negligible, so a demand-based APT justifies retaining the K dominant slopes and adding a single residual elasticity for idiosyncratic shocks, yielding a parsimonious $(K+1)$-parameter specification.

Start with the N-factor decomposition of trading demand $\Delta\mathbf{Q} = \sum_{k=1}^N \mathbf{w}_k \Delta Q_k^F$ from Proposition 2. Each factor is normalized to unit return variance, $\text{var}(\mathbf{w}_k^\top \mathbf{R}_0) = 1$, and the demand variances are ordered $\text{var}(\Delta Q_1^F) > \text{var}(\Delta Q_2^F) > \cdots > \text{var}(\Delta Q_N^F)$. The share of demand-induced risk explained by the top K factors is therefore proportional to $\sum_{k=1}^K \text{var}(\Delta Q_k^F)$. In practice a few factors often suffice; for example, An and Huber (2025) report that the three leading factors account for roughly 90% of demand-induced risk in trading among 17 major currencies.

Retaining only the leading factors yields the truncated model

$$\Delta\check{\mathbf{p}}(\Delta\mathbf{Q}) = \sum_{k=1}^K \lambda_k \Delta Q_k^F \text{cov}(\mathbf{R}_0, \mathbf{w}_k^\top \mathbf{R}_0), \tag{26}$$

to be compared with the full specification

$$\Delta\mathbf{p}(\Delta\mathbf{Q}) = \sum_{k=1}^{N} \lambda_k \Delta Q_k^F \operatorname{cov}(\mathbf{R}_0, \mathbf{w}_k^\top \mathbf{R}_0). \tag{27}$$

The next proposition gives conditions under which the pricing error $\Delta\mathbf{p}(\Delta\mathbf{Q}) - \Delta\check{\mathbf{p}}(\Delta\mathbf{Q})$ vanish. Appendix B.4 contains the proof.

PROPOSITION 3. *Suppose* $|\lambda_k| \leq H$ *for all* $k = 1, \ldots, N$ *for some constant* H. *If* $\sum_{k=K+1}^{N} \operatorname{var}(\Delta Q_k^F) \to 0$, *then* $\mathbb{E}\left[\|\Delta\mathbf{p}(\Delta\mathbf{Q}) - \Delta\check{\mathbf{p}}(\Delta\mathbf{Q})\|^2\right] \to 0$, *where* $\|\cdot\|$ *is the* L^2 *norm.*

Because each factor is scaled to unit return variance, $\operatorname{var}(\Delta Q_k^F)$ measures the demand-induced risk for that factor. When the total risk of the tail factors becomes negligible, the pricing error from dropping them also disappears—provided the price sensitivity to risk, λ_k, is uniformly bounded. The bound on λ_k mirrors the "good-deal bound" of Cochrane and Saa-Requejo (2000): if any factor offered an outsized marginal risk-return trade-off, arbitrage would quickly eliminate it.

The residual demand carried by the last $N - K$ portfolios can be written as

$$\mathbf{e} := \sum_{k=K+1}^{N} \mathbf{w}_k \Delta Q_k^F, \tag{28}$$

where the n-th element e_n can be interpreted as the idiosyncratic demand for asset n. Proposition 3 shows that the price impact of $\mathbf{e}$ is small but does not pin down its exact form. A pragmatic approach is to treat these residual shocks asset-by-asset. Because they are, by construction, small and largely uncorrelated, their cross-impacts are hard to detect empirically. Assume therefore that each e_n shifts only its own price and estimate a common

inverse-demand elasticity

$$\theta := \frac{\Delta P_n / P_n}{e_n / Q_n}, \tag{29}$$

an assumption supported by the approximately constant elasticities across assets documented in Davis (2025). Augmenting the K-factor model (26) yields

$$\Delta\tilde{\mathbf{p}}(\Delta\mathbf{Q}) = \sum_{k=1}^{K} \lambda_k \Delta Q_k^F \operatorname{cov}(\mathbf{R}_0, \mathbf{w}_k^\top \mathbf{R}_0) + \theta\, \mathbf{e}/\mathbf{Q}, \tag{30}$$

where $\mathbf{e}/\mathbf{Q}$ denotes element-wise division. The resulting specification has K factor slopes $\lambda_1, \lambda_2, \ldots, \lambda_K$ plus a single residual elasticity θ.

4.3 Why a K^2 Truncation Fails

Sections 3.3-4.2 reduced dimensionality in two stages,

$$N^2 \xrightarrow{\text{IUP}} N \xrightarrow{\text{demand-based APT}} K.$$

A tempting shortcut is to skip IUP and shrink the full N^2 model

$$\Delta\mathbf{p}(\Delta\mathbf{Q}) = \sum_{k=1}^{N} \sum_{l=1}^{N} \lambda_{k,l} \Delta Q_l^F \operatorname{cov}(\mathbf{R}_0, \mathbf{w}_k^\top \mathbf{R}_0) \tag{31}$$

directly to a K^2 version that retains only the first K factors,

$$\Delta\mathbf{p}_K(\Delta\mathbf{Q}) = \sum_{k=1}^{K} \sum_{l=1}^{K} \lambda_{k,l} \Delta Q_l^F \operatorname{cov}(\mathbf{R}_0, \mathbf{w}_k^\top \mathbf{R}_0). \tag{32}$$

Here every retained factor may still influence every other retained factor, so K^2 free slopes remain. The question is whether $\Delta\mathbf{p}_K(\Delta\mathbf{Q})$ can approximate the full impact map $\Delta\mathbf{p}(\Delta\mathbf{Q})$

on theoretically sound grounds.

Write the approximation error as

$$\Delta\mathbf{p}(\Delta\mathbf{Q})-\Delta\mathbf{p}_K(\Delta\mathbf{Q}) = \underbrace{\sum_{k=1}^{N}\sum_{l=K+1}^{N}\lambda_{k,l}\Delta Q_l^F \mathrm{cov}(\mathbf{R}_0, \mathbf{w}_k^\top \mathbf{R}_0)}_{\text{(A): discarded-demand block}} + \underbrace{\sum_{k=K+1}^{N}\sum_{l=1}^{K}\lambda_{k,l}\Delta Q_l^F \mathrm{cov}(\mathbf{R}_0, \mathbf{w}_k^\top \mathbf{R}_0)}_{\text{(B): spillover-to-discarded block}}. \tag{33}$$

Block (A) captures the effect of the discarded demands $\Delta Q_{K+1}^F, \ldots, \Delta Q_N^F$ on all prices. Under the conditions of Proposition 3—uniformly bounded slopes and vanishing variances $\mathrm{var}(\Delta Q_{K+1}^F), \ldots, \mathrm{var}(\Delta Q_N^F)$—this block indeed converges to zero.

Block (B) is the sticking point. It captures how the *retained* demands $\Delta Q_1^F, \ldots, \Delta Q_K^F$ spill into the prices of the *discarded* portfolios $\mathbf{w}_{K+1}, \ldots, \mathbf{w}_N$. Because the variances of $\Delta Q_1^F, \ldots, \Delta Q_K^F$ do not vanish, the only way to eliminate Block (B) is to assume—by fiat—that the corresponding slopes $\lambda_{k,l}$ are exactly zero. Such ad-hoc zeroing lacks economic justification: before orthogonalization, factor shocks can in principle affect any portfolio. In equity markets, for example, the boundary between "factor" and "idiosyncratic" risk is already fuzzy (see Cochrane, 2011 on the "factor zoo"), so declaring cross-impact to vanish precisely at K would be arbitrary.

The two-step path developed here avoids that pitfall. First, IUP diagonalizes the cross-impact matrix, reducing the parameter count from N^2 to N and ensuring that each canonical factor moves only its own price of risk. Second, when the top K factors explain the majority of demand-induced risk, a demand-based APT shows that the last $N-K$ factors contribute negligibly to prices, shrinking the system from N parameters to K without any ad-hoc restrictions. A direct K^2 truncation therefore lacks theoretical support, whereas the IUP-plus-APT route provides a coherent and empirically viable reduction.

5 Robustness and Theoretical Extensions

The analysis so far relies on two modeling choices: the canonical IUP factors and the linear price-impact term $\lambda_k \Delta Q_k^F$. This section shows that the main results do not depend on either choice. Section 5.1 proves that, among all ways to select N portfolios and eliminate cross-impact, the IUP basis introduces the least distortion when moving from the full N^2 specification to the N-factor representation. Section 5.2 replaces the linear term with a flexible response curve $g_k(\Delta Q_k^F)$ for each factor; because IUP still blocks cross-factor spillovers, each curve operates on its factor alone. Together, these extensions confirm that the factor structure—and the demand-based APT built on it—is a robust implication of the axioms, not an artifact of the particular factor choice or the assumption of linear price impact.

5.1 Optimality of the Canonical Factors

IUP eliminates cross-impacts between portfolios whose trading demands and fundamental returns are jointly uncorrelated—a restriction that empirical work supports but that also has a deeper theoretical basis. Any attempt to compress an N^2 system to N parameters must discard information, so the choice of basis matters. Under regularity Assumption 5 (with the full treatment in Online Appendix A), this subsection shows that the IUP basis is unique in preserving two key objects: (i) each portfolio's price sensitivity to its own demand and (ii) the marginal investor's expected utility. No other selection of N portfolios that removes cross-impact can satisfy both properties simultaneously, establishing the IUP factors as the least-distorting way to move from the full N^2 representation to the N-parameter factor form.

To formalize the concept, begin with the full N^2 model (10) and define a "model orthogonalization" as any rotation that selects N linearly independent portfolios and zeroes out

their cross-impacts.

DEFINITION 1. *Let $\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots, \tilde{\mathbf{w}}_N$ be N linearly independent portfolios. Expressing the full model in this basis gives*

$$\Delta\mathbf{p}(\Delta\mathbf{Q}) = \sum_{k=1}^{N}\sum_{l=1}^{N} \tilde{\lambda}_{k,l}\Delta\tilde{Q}_l^F \mathrm{cov}(\mathbf{R}_0, \tilde{\mathbf{w}}_k^\top \mathbf{R}_0), \tag{34}$$

where $\tilde{\lambda}_{k,l}$ are the N^2 free slope parameters and $\Delta\tilde{Q}_l^F$ are the trading demands for the rotated portfolios. The orthogonalized N-factor model associated with this basis sets the off-diagonal slopes to zero:

$$\Delta\bar{\mathbf{p}}(\Delta\mathbf{Q}) = \sum_{k=1}^{N} \tilde{\lambda}_{k,k}\Delta\tilde{Q}_k^F \mathrm{cov}(\mathbf{R}_0, \tilde{\mathbf{w}}_k^\top \mathbf{R}_0). \tag{35}$$

By Theorem 1, the IUP basis is one particular orthogonalization: it selects the portfolios $\mathbf{w}_1, \ldots, \mathbf{w}_N$ with uncorrelated fundamental returns and uncorrelated demands as the canonical factors. The following paragraphs provide quick intuition for why these two requirements are economically important. A formal statement appears in Theorem 2.

First, consider how a one-unit shock to factor-k demand affects its own price. In the full N^2 model (34) the sensitivity is

$$\frac{\partial\tilde{\mathbf{w}}_k^\top\Delta\mathbf{p}(\Delta\mathbf{Q})}{\partial\Delta\tilde{Q}_k^F} = \tilde{\lambda}_{k,k}\mathrm{var}(\tilde{\mathbf{w}}_k^\top\mathbf{R}_0) + \sum_{l\neq k}\tilde{\lambda}_{l,k}\mathrm{cov}(\tilde{\mathbf{w}}_k^\top\mathbf{R}_0, \tilde{\mathbf{w}}_l^\top\mathbf{R}_0). \tag{36}$$

The first term is the direct effect; the second term captures an indirect channel. Demand for factor k also moves the price of factor l through the cross-slope $\tilde{\lambda}_{l,k}$. Because factors k and l have correlated fundamental returns, that price change arbitrages back into factor k, adding the covariance-weighted spillover shown above.

In the orthogonalized N-factor model (35) all off-diagonal slopes are set to zero, so the

sensitivity reduces to

$$\frac{\partial \tilde{\mathbf{w}}_k^\top \Delta \bar{\mathbf{p}}(\Delta \mathbf{Q})}{\partial \Delta \tilde{Q}_k^F} = \tilde{\lambda}_{k,k} \text{var}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0). \tag{37}$$

Thus, if the factor basis is chosen so that fundamental returns are uncorrelated—$\text{cov}(\tilde{\mathbf{w}}_l^\top \mathbf{R}_0, \tilde{\mathbf{w}}_k^\top \mathbf{R}_0) = 0$ for any $l \neq k$—the indirect spillover vanishes and each factor's price sensitivity to its own demand is preserved when moving from the full N^2 model to the N-factor representation. Empirically, this ensures that regressing a factor's price change on its own demand recovers the slope $\tilde{\lambda}_{k,k}$ without bias.

Second, consider the marginal investor's expected utility. Let $W(\Delta \mathbf{Q})$ be the end-of-period wealth after absorbing the demand $\Delta \mathbf{Q}$, and let $u(\cdot)$ denote the utility function. With the demand absorbed, expected utility is

$$\mathbb{E}[u(W(\Delta \mathbf{Q}))] = \mathbb{E}[u(W(\mathbf{0}) + R_F \Delta \mathbf{Q}^\top \Delta \mathbf{p}(\Delta \mathbf{Q}) - \Delta \mathbf{Q}^\top (\mathbf{R}_0 - R_F \mathbf{1}))], \tag{38}$$

where $\Delta \mathbf{Q}^\top (\mathbf{R}_0 - R_F \mathbf{1})$ is the time-1 payoff change and $\Delta \mathbf{Q}^\top \Delta \mathbf{p}(\Delta \mathbf{Q})$ is the time-0 cash compensation for bearing that risk.

Because $\Delta \mathbf{Q}$ is independent of the payoff $\mathbf{X}$, the compensation $\Delta \mathbf{Q}^\top \Delta \mathbf{p}(\Delta \mathbf{Q})$ is independent of $\mathbf{X}$ as well. To isolate utility curvature to fundamental payoff risk, assume the investor is risk-averse only to $\mathbf{X}$. Formally, there exists a function $\tilde{u}(\cdot)$ such that

$$u(W(\mathbf{0}) + R_F \Delta \mathbf{Q}^\top \Delta \mathbf{p}(\Delta \mathbf{Q}) - \Delta \mathbf{Q}^\top (\mathbf{R}_0 - R_F \mathbf{1})) = R_F \Delta \mathbf{Q}^\top \Delta \mathbf{p}(\Delta \mathbf{Q}) + \tilde{u}(W(\mathbf{0}) - \Delta \mathbf{Q}^\top (\mathbf{R}_0 - R_F \mathbf{1})). \tag{39}$$

Substituting (39) into (38) yields

$$\mathbb{E}[u(W(\Delta \mathbf{Q}))] = R_F \mathbb{E}[\Delta \mathbf{Q}^\top \Delta \mathbf{p}(\Delta \mathbf{Q})] + \mathbb{E}[\tilde{u}(W(\mathbf{0}) - \Delta \mathbf{Q}^\top (\mathbf{R}_0 - R_F \mathbf{1}))]. \tag{40}$$

Thus price impact influences expected utility solely through the expected compensation $\mathbb{E}[\Delta\mathbf{Q}^\top \Delta\mathbf{p}(\Delta\mathbf{Q})]$. Any orthogonalization that leaves this expectation unchanged preserves the marginal investor's expected utility, regardless of the utility function.

Under the full N^2 model (34) the expected compensation is

$$\mathbb{E}[\Delta\mathbf{Q}^\top \Delta\mathbf{p}(\Delta\mathbf{Q})] = \sum_{k=1}^{N} \tilde{\lambda}_{k,k} \mathrm{var}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0) \mathrm{var}\left(\Delta\tilde{Q}_k^F\right) + \sum_{\substack{k,l=1 \\ k\neq l}}^{N} \tilde{\lambda}_{k,l} \mathrm{var}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0) \mathrm{cov}\left(\Delta\tilde{Q}_k^F, \Delta\tilde{Q}_l^F\right), \tag{41}$$

which uses $\mathrm{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_l^\top \mathbf{R}_0) = 0$ for $k \neq l$ and $\Delta\mathbf{Q}$ is mean-zero. In the orthogonalized N-factor model (35) all off-diagonal slopes are set to zero, so the expected compensation becomes

$$\mathbb{E}[\Delta\mathbf{Q}^\top \Delta\bar{\mathbf{p}}(\Delta\mathbf{Q})] = \sum_{k=1}^{N} \tilde{\lambda}_{k,k} \mathrm{var}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0) \mathrm{var}\left(\Delta\tilde{Q}_k^F\right). \tag{42}$$

The two expressions coincide precisely when the factor demands are mutually uncorrelated, $\mathrm{cov}\left(\Delta\tilde{Q}_k^F, \Delta\tilde{Q}_l^F\right) = 0$ for any $k \neq l$. Choosing a basis with uncorrelated demands therefore guarantees that expected compensation—and hence expected utility—remains unchanged when moving from the full N^2 model (34) to the N-factor representation (35).

Theorem 2 formalizes the foregoing intuition; its proof appears in Appendix B.5. The theorem shows that the IUP basis is the *only* orthogonalization that leaves both each factor's own price sensitivity and the investor's expected utility unchanged, establishing its optimality in a necessary-and-sufficient sense.

THEOREM 2. *Fix any orthogonalization $\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots, \tilde{\mathbf{w}}_N$. Consider the full model (34) with slopes $\tilde{\lambda}_{k,l}$ for $k, l = 1, 2, \ldots, N$ and its orthogonalized counterpart (35) with slopes $\tilde{\lambda}_{1,1}, \ldots, \tilde{\lambda}_{N,N}$. The orthogonalized model satisfies IUP for every choice of $\tilde{\lambda}_{k,k}$ if and only if the following two conditions hold for every set of $\tilde{\lambda}_{k,l}$:*

- *Invariance of own-price sensitivity: for each k,*

$$\frac{\partial \tilde{\mathbf{w}}_k^\top \Delta \mathbf{p}(\Delta \mathbf{Q})}{\partial \Delta \tilde{Q}_k^F} = \frac{\partial \tilde{\mathbf{w}}_k^\top \Delta \bar{\mathbf{p}}(\Delta \mathbf{Q})}{\partial \Delta \tilde{Q}_k^F}. \tag{43}$$

- *The marginal investor's expected utility $\mathbb{E}[u(W(\Delta \mathbf{Q}))]$ is the same under (34) and (35).*

5.2 Non-Linear Price-Impact Curves

Empirical evidence suggests that price impact becomes strongly concave once demand shock is large, implying that the true impact map is non-linear (Hasbrouck, 1991; Fan, Feng, Au, and Baronyan, 2022; Chaudhry and Li, 2025). Until now the analysis has relied on the linear price-impact axiom in Assumption 1. This subsection relaxes that axiom and develops a non-linear specification in which each factor has its own response curve.

Formally, replace the linear term with factor-specific response functions

$$\Delta \mathbf{p}(\Delta \mathbf{Q}) = \sum_{k=1}^{N} g_k(\Delta Q_k^F) \mathrm{cov}(\mathbf{R}_0, \mathbf{w}_k^\top \mathbf{R}_0), \tag{44}$$

where each $g_k(\cdot)$ maps demand for factor k into its own price change and satisfies $g_k(0) = 0$. The linear model (20) is recovered by setting $g_k(\Delta Q_k^F) = \lambda_k \Delta Q_k^F$; in general, the functions $g_k(\cdot)$ can be flexibly estimated from the data.

Although price impact is now nonlinear, the accounting identity for factor demand $\Delta \mathbf{Q} = \sum_{k=1}^{N} \mathbf{w}_k \Delta Q_k^F$ remains linear by construction. Moreover, the canonical factors of Proposition 2 still exhibit zero cross-impact, so each curve $g_k(\cdot)$ operates on its factor in isolation. The covariance term in (44) continues to enforce the cross-sectional no-arbitrage link.

6 Conclusion

In a market with N assets, a shock to the demand for one asset can, in principle, influence the prices of the other $N-1$. The reduced-form description therefore contains N^2 cross-impact slopes that record "who moves whom." Which of those N^2 interactions are pinned down by economic first principles, and which are free to reflect investor behavior? The axiomatic framework developed here answers that question without committing to any particular equilibrium. Starting from the most general no-arbitrage N^2-parameter representation, the analysis introduces a small set of transparent axioms and records the exact dimensionality reduction each produces. The *Irrelevance of Uncorrelated Portfolios* (IUP) axiom is the pivotal step: it is both necessary and sufficient to isolate N canonical factors that are mutually uncorrelated in fundamental risk and trading demand, thereby proving that each factor's demand moves only its own price of risk. The result is an N-parameter factor system whose slopes λ_k measure the marginal price of factor risk—conceptually distinct from both unconditional prices of risk and inverse-demand elasticities.

Once the IUP step has diagonalized cross-impact, dimensionality can fall along two independent routes, as depicted in Figure 1. Imposing an equal-slope condition across all factors condenses the system to the familiar one-parameter mean–variance formula. Alternatively, a demand-side extension of the Arbitrage Pricing Theory shows that when a small set of factors explains nearly all demand-induced risk, and the λ_k remain bounded, the model trims to a parsimonious $(K+1)$-parameter specification: the K dominant factor slopes plus a single residual elasticity that captures the effect of small, largely uncorrelated shocks. The two paths are complementary rather than sequential: one recovers the classical benchmark, the other yields a flexible yet tractable form well suited to empirical work.

Future work could enrich the framework by introducing heterogeneous investor types, allowing factor slopes to evolve with market conditions, or applying the methodology to markets beyond equities and foreign exchange. Such extensions would deepen the understanding of when, how, and by how much uninformed demand moves prices, while retaining the clarity afforded by the axiomatic approach developed here.

References

An, Yu, and Amy W. Huber, 2025, Demand propagation through traded risk factors, Working paper.

An, Yu, Yinan Su, and Chen Wang, 2025, Quantity, risk, and return, Working paper.

An, Yu, and Zeyu Zheng, 2025, Information from uninformed demand: How predictable demand propagates across assets, Working paper.

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

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

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

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

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

Chaudhry, Aditya, and Jiacui Li, 2025, Endogenous elasticities: Price multipliers are smaller for larger demand shocks, Working paper.

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

Cochrane, John H, 2011, Presidential address: Discount rates, *The Journal of finance* 66, 1047–1108.

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.

Davis, Carter, 2025, Classical versus demand system asset pricing, Working paper.

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.

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.

Fan, Yinghua, Guanhao Feng, Scott Au, and Sayad Baronyan, 2022, Estimating asymmetric price impact, Working paper.

Fuchs, William, Satoshi Fukuda, and Daniel Neuhann, 2025a, Demand-system asset pricing: Theoretical foundations, Working paper.

Fuchs, William, Satoshi Fukuda, and Daniel Neuhann, 2025b, A trilemma for asset demand estimation, Working paper.

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

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

Haddad, Valentin, Zhiguo He, Paul Huebner, Péter Kondor, and Erik Loualiche, 2025, Causal inference for asset pricing, Working Paper.

Hansen, Lars Peter, and Scott F Richard, 1987, The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models, *Econometrica* 587–613.

Hartzmark, Samuel M, and David H Solomon, 2025, Marketwide predictable price pressure, *American Economic Review* Forthcoming.

Hasbrouck, Joel, 1991, Measuring the information content of stock trades, *Journal of Finance* 46, 179–207.

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

Jansen, Kristy AE, Wenhao Li, and Lukas Schmid, 2025, Granular treasury demand with arbitrageurs, Working paper.

Jolliffe, Ian, 1986, *Principal Component Analysis* (Springer Series in Statistics).

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

Koijen, Ralph SJ, and Motohiro Yogo, 2025, On the theory and econometrics of (demand system) asset pricing, *Available at SSRN 5274709* .

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.

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

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

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, 2023, Imperfect competition in financial markets: Recent developments, Working paper.

Sharpe, William F, 1964, Capital asset prices: A theory of market equilibrium under conditions of risk, *The Journal of Finance* 19, 425–442.

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

Vayanos, Dimitri, and Jean-Luc Vila, 2021, A preferred-habitat model of the term structure of interest rates, *Econometrica* 89, 77–112.

Appendix

The appendices provide additional theoretical results and proofs.

A SDF Change and Price Impact Under Alternative Factors

Empirical applications often work with pre-defined portfolios—such as the Fama–French factors—that are not necessarily orthogonal in either returns or demands. Using such portfolios re-introduces an $N \times N$ cross-impact matrix, because the factors generally influence one another. Even so, the IUP axiom keeps the representation parsimonious: for any such factor basis the cross-impact matrix is still pinned down by only N free parameters, not N^2. This appendix spells out the resulting structural restrictions.

Let $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_N)$ be an arbitrary invertible matrix of portfolio weights. For illustration, $\tilde{\mathbf{w}}_1$ could be the Fama-French high-minus-low (HML) portfolio and $\tilde{\mathbf{w}}_2$ the small-minus-big (SMB) portfolio. Define factor-level demands by $\Delta\mathbf{Q} = \tilde{\mathbf{W}}\Delta\tilde{\mathbf{Q}}^F$, $\Delta\tilde{\mathbf{Q}}^F = (\Delta\tilde{Q}_1^F, \ldots, \Delta\tilde{Q}_N^F)^\top$, so that $\Delta\tilde{Q}_1^F$ is the demand for HML, $\Delta\tilde{Q}_2^F$ for SMB, and so on. These factors are typically correlated, so cross-impact terms appear. Even with such a basis, IUP restricts the slope matrix to N degrees of freedom.

PROPOSITION 4. *Under alternative factors, the SDF change in Theorem 1 can be written*

$$\Delta M(\Delta\mathbf{Q}) = \sum_{k=1}^{N}\sum_{j=1}^{N} \tilde{\lambda}_{k,j}\Delta\tilde{Q}_j^F \tilde{\mathbf{w}}_k^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) = (\tilde{\mathbf{\Lambda}}\Delta\tilde{\mathbf{Q}}^F)^\top\tilde{\mathbf{W}}^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]), \tag{A.1}$$

where $\tilde{\mathbf{\Lambda}} = \{\tilde{\lambda}_{k,j}\}$ *is an* $N \times N$ *slope matrix. The corresponding price-impact map is*

$$\Delta\mathbf{p}(\Delta\mathbf{Q}) = \sum_{k=1}^{N}\sum_{j=1}^{N} \tilde{\lambda}_{k,j}\Delta\tilde{Q}_j^F \operatorname{cov}(\mathbf{R}_0, \tilde{\mathbf{w}}_k^\top\mathbf{R}_0). \tag{A.2}$$

The matrix $\tilde{\mathbf{\Lambda}}$ satisfies, for some invertible matrix $\mathbf{O}$:

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

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

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

Proof. Because $\tilde{\mathbf{W}}$ and $\mathbf{W}$ each span $\mathbb{R}^N$, there exists an invertible matrix $\mathbf{O}$ such that $\tilde{\mathbf{W}}\mathbf{O} = \mathbf{W}$ and $\mathbf{O}\Delta\mathbf{Q}^F = \Delta\tilde{\mathbf{Q}}^F$, where $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_N)$ and $\Delta\mathbf{Q}^F = (\Delta Q_1^F, \ldots, \Delta Q_N^F)^{\top}$ are constructed in Proposition 2. Equation (A.1) implies that

$$\Delta M(\Delta\mathbf{Q}) = (\mathbf{O}^{-1}\tilde{\mathbf{\Lambda}}\mathbf{O}\Delta\mathbf{Q}^F)^{\top}\mathbf{W}^{\top}(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]). \tag{A.6}$$

By (19), the above equation implies that $\mathbf{O}^{-1}\tilde{\mathbf{\Lambda}}\mathbf{O} = \mathbf{\Lambda}$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ is a diagonal matrix. Since $\mathbf{W}$ and $\Delta\mathbf{Q}^F$ satisfy Proposition 2, conditions (17) and (18) translate into (A.4) and (A.5). □

B Proofs

This appendix provides proofs omitted in the main text.

B.1 Proof of Proposition 1

First, let the projection of $\Delta M(\Delta\mathbf{Q})$ onto the space spanned by $\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]$ be denoted as $\Delta M(\Delta\mathbf{Q}) = (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])^{\top}\mathbf{b}$ for some $\mathbf{b} \in \mathbb{R}^N$. Equation (5) implies that $\Delta\mathbf{p}(\Delta\mathbf{Q}) = \mathrm{var}(\mathbf{R}_0)\mathbf{b}$. Therefore, $\mathbf{b} = \mathrm{var}(\mathbf{R}_0)^{-1}\Delta\mathbf{p}(\Delta\mathbf{Q})$, which gives

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

Next, equation (B.1) and the linearity assumption 1 imply that, for any $a_1 \in \mathbb{R}$, $a_2 \in \mathbb{R}$, $\Delta\mathbf{Q}_1 \in \mathbb{R}^N$, and $\Delta\mathbf{Q}_2 \in \mathbb{R}^N$,

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

Then, let $\mathbf{c}_n$ be an $N\times 1$ vector with the n-th entry equal to one and all other entries equal to zero. Equation (B.2) implies that $\Delta M(\Delta\mathbf{Q}) = \sum_{n=1}^N \Delta Q_n \Delta M(\mathbf{c}_n)$. Because $\Delta M(\mathbf{c}_n)$ also lies within the space spanned by $\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]$, one can write $\Delta M(\mathbf{c}_n) = \mathbf{y}_n^\top(\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0])$ for some $\mathbf{y}_n \in \mathbb{R}^N$. Therefore,

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

where $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)$. Finally, equation (10) and Assumption 2 imply that $\mathrm{var}(\mathbf{R}_0)\mathbf{Y}$ is positive definite.

Conversely, it is straightforward to see that $\Delta M(\Delta\mathbf{Q})$ in (9) satisfies Assumptions 1, 2, and 3.

B.2 Proof of Proposition 2

Since the matrix $\mathrm{var}(\mathbf{R}_0)$ has full rank, a Cholesky decomposition yields $\mathrm{var}(\mathbf{R}_0) = \mathbf{U}^\top\mathbf{U}$, where $\mathbf{U}$ is an $N \times N$ upper triangular matrix with positive diagonal entries. Next, perform an eigenvalue decomposition of the symmetric matrix $\mathbf{U}\mathrm{var}(\Delta\mathbf{Q})\mathbf{U}^\top$. Since $\mathrm{var}(\Delta\mathbf{Q})$ has

full rank, this gives:

$$\mathbf{U}\text{var}(\Delta\mathbf{Q})\mathbf{U}^\top\mathbf{G} = \mathbf{G}\mathbf{\Pi}, \tag{B.4}$$

where $\mathbf{\Pi} = \text{diag}(\pi_1, \ldots, \pi_N)$ and $\mathbf{G}$ is an $N \times N$ orthonormal matrix satisfying $\mathbf{G}^\top\mathbf{G} = \mathbf{I}_N$. Set $\mathbf{W} = \mathbf{U}^{-1}\mathbf{G}$; $\mathbf{U}$ is invertible because it is upper-triangular with positive diagonal entries.

The matrix $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_N)$ constructed above satisfies the conditions (17) and (18). First,

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

Second, equation (16) implies $\Delta\mathbf{Q} = \mathbf{W}\Delta\mathbf{Q}^F$, where $\Delta\mathbf{Q}^F = (\Delta Q_1^F, \Delta Q_2^F, \ldots, \Delta Q_N^F)^\top$. Therefore,

$$\text{var}(\Delta\mathbf{Q}) = \mathbf{W}\text{var}(\Delta\mathbf{Q}^F)\mathbf{W}^\top. \tag{B.6}$$

By (B.4), $\mathbf{U}\text{var}(\Delta\mathbf{Q})\mathbf{U}^\top\mathbf{U}\mathbf{W} = \mathbf{U}\mathbf{W}\mathbf{\Pi}$. Because $\mathbf{U}$ is invertible,

$$\text{var}(\Delta\mathbf{Q})\mathbf{U}^\top\mathbf{U}\mathbf{W} = \mathbf{W}\mathbf{\Pi}. \tag{B.7}$$

Substituting (B.6) into (B.7),

$$\mathbf{W}\mathbf{\Pi} = \mathbf{W}\text{var}(\Delta\mathbf{Q}^F)\mathbf{W}^\top\mathbf{U}^\top\mathbf{U}\mathbf{W} = \mathbf{W}\text{var}(\Delta\mathbf{Q}^F), \tag{B.8}$$

because $\mathbf{W}^\top\mathbf{U}^\top\mathbf{U}\mathbf{W} = \mathbf{G}^\top\mathbf{G} = \mathbf{I}_N$. Equation (B.8) implies $\text{var}(\Delta\mathbf{Q}^F) = \mathbf{\Pi}$.

Finally, uniqueness of the matrix $\mathbf{W}$ is demonstrated. Suppose another matrix $\tilde{\mathbf{W}}$ also satisfies (17) and (18). Define $\tilde{\mathbf{G}} = \mathbf{U}\tilde{\mathbf{W}}$. This yields $\tilde{\mathbf{G}}^\top\tilde{\mathbf{G}} = \mathbf{I}_N$ and $\tilde{\mathbf{G}}^\top\mathbf{U}\text{var}(\Delta\mathbf{Q})\mathbf{U}^\top\tilde{\mathbf{G}} = \mathbf{\Pi}$. Assumption 5 requires that $\mathbf{U}\text{var}(\Delta\mathbf{Q})\mathbf{U}^\top$ has distinct eigenvalues, implying that the matrix $\tilde{\mathbf{G}}$ is unique up to sign for any column (since the eigenvalue matrix $\mathbf{\Pi}$ is ordered

from largest to smallest eigenvalue). Note that $\tilde{\mathbf{W}} = \mathbf{U}^{-1}\tilde{\mathbf{G}}$, where $\mathbf{U}^{-1}$ is also an upper triangular matrix. Thus, the matrix $\tilde{\mathbf{W}}$ is also unique up to sign for any column.

B.3 Proof of Theorem 1

Begin by establishing the following lemma, which allows one to state the equivalent IUP condition under the factors $\mathbf{w}_1, \ldots, \mathbf{w}_N$.

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

Proof. Following the proof of Proposition 2 in Appendix B.2, one carries out Cholesky decomposition to obtain $\text{var}(\mathbf{R}_0) = \mathbf{U}^\top \mathbf{U}$, where $\mathbf{U}$ is an $N \times N$ upper triangular matrix. Then, $\mathbf{w}_k = \mathbf{U}^{-1}\mathbf{g}_k$ for $k = 1, 2, \ldots, N$, where $\mathbf{g}_1, \ldots, \mathbf{g}_N$ is the full set of eigenvectors of $\text{var}(\mathbf{U}\Delta\mathbf{Q})$ with distinct eigenvalues $\pi_1 > \pi_2 > \cdots > \pi_N > 0$.

Observe that

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

where (B.9) follows from the Cholesky decomposition of $\text{var}(\mathbf{R}_0)$.

Define the following set of matrices,

$$
\mathcal{H} = \left\{ \mathbf{H} \in \mathbb{R}^{N \times N} \,\middle|\, \mathbf{H} = \sum_{k=1}^{N} z_k \mathbf{g}_k \mathbf{g}_k^\top \text{ for some real numbers } z_1, z_2, \ldots, z_N \right\}. \qquad \text{(B.11)}
$$

The claim is that

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

If $\mathbf{H} \in \mathcal{H}$ as defined in (B.11), then $\mathbf{H}$ also lies in (B.12) because $\mathbf{g}_1, \ldots, \mathbf{g}_N$ are the eigenvectors of $\mathbf{U}\text{var}(\Delta\mathbf{Q})\mathbf{U}^\top$. Conversely, if $\mathbf{H} \in \mathcal{H}$ as defined in (B.12), then for any k,

$$\mathbf{U}\text{var}(\Delta\mathbf{Q})\mathbf{U}^\top(\mathbf{H}\mathbf{g}_k) = (\mathbf{U}\text{var}(\Delta\mathbf{Q})\mathbf{U}^\top\mathbf{H})\mathbf{g}_k = (\mathbf{H}\mathbf{U}\text{var}(\Delta\mathbf{Q})\mathbf{U}^\top)\mathbf{g}_k = \pi_k(\mathbf{H}\mathbf{g}_k), \tag{B.13}$$

showing that $\mathbf{H}\mathbf{g}_k$ is also the eigenvector of $\mathbf{U}\text{var}(\Delta\mathbf{Q})\mathbf{U}^\top$ that corresponds to the eigenvalue π_k. Because the eigenvalues are distinct and the eigenspaces are all one-dimensional, one has $\mathbf{H}\mathbf{g}_k = \mu_k\mathbf{g}_k$ for some μ_k. This shows that $\mathbf{g}_k$ is also an eigenvector of $\mathbf{H}$. Because k is arbitrary in $1, \ldots, N$, it follows that $\mathbf{H} = \sum_{k=1}^{N} \mu_k \mathbf{g}_k \mathbf{g}_k^\top$, so $\mathbf{H} \in \mathcal{H}$ of (B.11).

Given (B.12), equation (B.10) can be rewritten as

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

Define $\mathbf{C} = \mathbf{U}^\top\mathbf{H}\mathbf{U}$. Because $\mathbf{U}$ is invertible and $\text{var}(\mathbf{R}_0) = \mathbf{U}^\top\mathbf{U}$, $\mathbf{C} \in \mathcal{C} \iff \mathbf{H} \in \mathcal{H}$, where $\mathcal{C}$ is defined in (15). Therefore, one can rewrite (B.14) as $\mathbf{a}^\top\mathbf{C}\mathbf{d} = 0$ for any $\mathbf{C} \in \mathcal{C}$, which completes the proof of Lemma 1. □

Lemma 1 allows the IUP to be stated under the factor basis $\mathbf{w}_1, \ldots, \mathbf{w}_N$.

Assumption 4*. ***IUP under factors***

Let s be the trading-demand shock to a portfolio with weight vector

$$\mathbf{d} = \left(\text{cov}(\Delta Q_1, s), \text{cov}(\Delta Q_2, s), \ldots, \text{cov}(\Delta Q_N, s)\right)^\top, \tag{B.15}$$

and let $\mathbf{a} \in \mathbb{R}^N$ *be any portfolio weight vector. If* $\text{cov}(\mathbf{w}_k^\top \mathbf{R}_0, \mathbf{a}^\top \mathbf{R}_0)\text{cov}(\mathbf{w}_k^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) = 0$ *for all* k*, then* $\text{cov}(\mathbf{a}^\top \Delta\mathbf{p}(\Delta\mathbf{Q}), s) = 0$.

Note that the IUP condition in Assumption 4 is equivalent to Assumption 4* only when Assumption 5 holds. Without Assumption 5, the IUP imposes no restrictions for eigenvectors that correspond to the same eigenvalue. In that case, the factors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N$ are also not unique and the IUP condition does not simplify to Assumption 4*. Online Appendix A presents the general case.

One also has the following coordinate transformation,

$$\mathbf{d} = \text{cov}(\Delta\mathbf{Q}, s) = \sum_{k=1}^{N} \mathbf{w}_k \text{cov}(\Delta Q_k^F, s), \tag{B.16}$$

where the first equality follows from definition (B.15) and the second equality uses the factor decomposition (16). Using (B.16) and orthogonalization condition (17),

$$\text{cov}(\mathbf{w}_k^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) = \mathbf{w}_k^\top \text{var}(\mathbf{R}_0) \left(\sum_{l=1}^{N} \mathbf{w}_l \text{cov}(\Delta Q_l^F, s) \right) = \text{cov}(\Delta Q_k^F, s). \tag{B.17}$$

With these preliminaries, proceed to the theorem. First, show that the SDF change (19) satisfies the IUP in Assumption 4*. The form (19) is a special case of (9) and therefore automatically satisfies Assumptions 1 and 3. The price impact under the SDF change (19) is

$$\Delta\mathbf{p}(\Delta\mathbf{Q}) = \mathbb{E}[\mathbf{R}_0 \Delta M(\Delta\mathbf{Q})] = \text{var}(\mathbf{R}_0) \sum_{k=1}^{N} \lambda_k \Delta Q_k^F \mathbf{w}_k. \tag{B.18}$$

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

$$\mathbf{a} = \sum_{k=1}^{N} x_k \mathbf{w}_k \tag{B.19}$$

for some real numbers $x_1, x_2, \ldots, x_N$. The condition of the IUP in Assumption 4* is

$$\begin{aligned} 0 &= \text{cov}(\mathbf{w}_k^\top \mathbf{R}_0, \mathbf{a}^\top \mathbf{R}_0)\text{cov}(\mathbf{w}_k^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) \\ &= \left(\sum_{l=1}^{N} x_l \mathbf{w}_l\right)^\top \text{var}(\mathbf{R}_0)\mathbf{w}_k \text{cov}(\Delta Q_k^F, s) = x_k \text{cov}(\Delta Q_k^F, s), \end{aligned} \tag{B.20}$$

for any $k = 1, 2, \ldots, N$, which uses equations (17), (B.17), and (B.19).

Observe that

$$\mathbf{a}^\top \Delta \mathbf{p}(\Delta \mathbf{Q}) = \left(\sum_{l=1}^{N} x_l \mathbf{w}_l\right)^\top \text{var}(\mathbf{R}_0) \left(\sum_{k=1}^{N} \lambda_k \Delta Q_k^F \mathbf{w}_k\right) = \sum_{k=1}^{N} x_k \lambda_k \Delta Q_k^F, \tag{B.21}$$

which uses equations (17), (B.18), and (B.19). Therefore,

$$\text{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\Delta \mathbf{Q}), s) = \sum_{k=1}^{N} x_k \text{cov}(\Delta Q_k^F, s)\lambda_k. \tag{B.22}$$

Note that the IUP's assumption implies that $x_k \text{cov}(\Delta Q_k^F, s) = 0$ for any $k = 1, 2, \ldots, N$ in (B.20). Therefore, by equation (B.22), $\text{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\Delta \mathbf{Q}), s) = 0$, thereby showing that the IUP holds. Moreover, equation (B.18) implies that

$$\Delta \mathbf{Q}^\top \Delta \mathbf{p}(\Delta \mathbf{Q}) = (\Delta \mathbf{Q}^F)^\top \mathbf{W}^\top \text{var}(\mathbf{R}_0)\mathbf{W}\mathbf{\Lambda}\Delta \mathbf{Q}^F = (\Delta \mathbf{Q}^F)^\top \mathbf{\Lambda}\Delta \mathbf{Q}^F, \tag{B.23}$$

where $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)$, $\Delta \mathbf{Q}^F = (\Delta Q_1^F, \Delta Q_2^F, \ldots, \Delta Q_N^F)^\top$, and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$. Because $\mathbf{\Lambda}$ is positive definite, Assumption 2 holds.

Second, show that, if the SDF change (9) satisfies the IUP, it can be written in the form

(19). Using equation (16), equation (9) can be simplified as

$$\Delta M(\Delta \mathbf{Q}) = \sum_{k=1}^{N} \Delta Q_k^F (\mathbf{Y}\mathbf{w}_k)^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) = \sum_{k=1}^{N} \Delta Q_k^F \mathbf{h}_k^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]), \tag{B.24}$$

where $\mathbf{h}_k := \mathbf{Y}\mathbf{w}_k$. Therefore, the price impact under (B.24) is

$$\Delta \mathbf{p}(\Delta \mathbf{Q}) = \mathbb{E}[\mathbf{R}_0 \Delta M(\Delta \mathbf{Q})] = \mathrm{var}(\mathbf{R}_0) \sum_{k=1}^{N} \Delta Q_k^F \mathbf{h}_k. \tag{B.25}$$

Apply the IUP with portfolio $\mathbf{a} = \mathbf{w}_l$ and $s = \Delta Q_m^F$ for any $l \neq m$. From (B.16) and orthogonalization condition (18),

$$\mathbf{d} = \sum_{k=1}^{N} \mathbf{w}_k \mathrm{cov}(\Delta Q_k^F, \Delta Q_m^F) = \pi_m \mathbf{w}_m. \tag{B.26}$$

Thus, $\mathrm{cov}(\mathbf{w}_k^\top \mathbf{R}_0, \mathbf{a}^\top \mathbf{R}_0)\mathrm{cov}(\mathbf{w}_k^\top \mathbf{R}_0, \mathbf{d}^\top \mathbf{R}_0) = 0$ for any $k = 1, 2, \ldots, N$, because of orthogonalization condition (17). Thus, the IUP implies that

$$0 = \mathrm{cov}(\mathbf{a}^\top \Delta \mathbf{p}(\Delta \mathbf{Q}), s) = \mathrm{cov}\left(\mathbf{w}_l^\top \mathrm{var}(\mathbf{R}_0) \sum_{k=1}^{N} \Delta Q_k^F \mathbf{h}_k, \Delta Q_m^F \right) = \pi_m \mathbf{w}_l^\top \mathrm{var}(\mathbf{R}_0) \mathbf{h}_m. \tag{B.27}$$

Thus, for any $l \neq m$,

$$\mathbf{w}_l^\top \mathrm{var}(\mathbf{R}_0) \mathbf{h}_m = 0. \tag{B.28}$$

Project $\mathbf{h}_m$ onto the factor portfolios $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N$ and obtain $\mathbf{h}_m = \sum_{k=1}^{N} \theta_{m,k} \mathbf{w}_k$. By the IUP condition (B.28) and orthogonalization condition (17), $\theta_{m,k} = 0$ for any $m \neq k$. Hence

the SDF change (B.24) can be rewritten as

$$\Delta M(\Delta \mathbf{Q}) = \sum_{k=1}^{N} \theta_{k,k} \Delta Q_k^F \mathbf{w}_k^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]). \tag{B.29}$$

Renaming $\theta_{k,k}$ as λ_k yields the form (19). Lastly, (B.23) and Assumption 2 implies that $\lambda_k > 0$ for all k. The proof is complete.

B.4 Proof of Proposition 3

The pricing error of the price impact simplifies to

$$\Delta \mathbf{p}(\Delta \mathbf{Q}) - \Delta \check{\mathbf{p}}(\Delta \mathbf{Q}) = \sum_{k=K+1}^{N} \lambda_k \Delta Q_k^F \mathrm{cov}(\mathbf{R}_0, \mathbf{w}_k^\top \mathbf{R}_0) = \sum_{k=K+1}^{N} \mathbf{c}_k \Delta Q_k^F, \tag{B.30}$$

with the $N \times 1$ vector $\mathbf{c}_k = \lambda_k \mathrm{cov}(\mathbf{R}_0, \mathbf{w}_k^\top \mathbf{R}_0)$. Because λ_k are uniformly bounded, each element of every $\mathbf{c}_k = (c_{1,k}, \ldots, c_{N,k})^\top$ is also uniformly bounded. Using the fact that ΔQ_k^F are uncorrelated and mean-zero,

$$\mathbb{E}[\|\Delta \mathbf{p}(\Delta \mathbf{Q}) - \Delta \check{\mathbf{p}}(\Delta \mathbf{Q})\|^2] = \sum_{n=1}^{N} \mathbb{E}\left(\sum_{k=K+1}^{N} c_{n,k} \Delta Q_k^F \right)^2 = \sum_{n=1}^{N} \sum_{k=K+1}^{N} c_{n,k}^2 \mathrm{var}(\Delta Q_k^F), \tag{B.31}$$

which tends to zero as $\sum_{k=K+1}^{N} \mathrm{var}(\Delta Q_k^F)$ tends to zero.

B.5 Proof of Theorem 2

First, the sufficiency direction is established. From Theorem 1 and Proposition 4, if the orthogonalized model (35) satisfies the IUP, there exists some $N \times N$ invertible matrix $\mathbf{O}$, such that $\mathbf{O}^{-1}\mathrm{diag}(\tilde{\lambda}_{1,1}, \tilde{\lambda}_{2,2}, \ldots, \tilde{\lambda}_{N,N})\mathbf{O}$ 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 $y_{n,m}$. Then, $x_{n,m}y_{m,l} = 0$ for any m and $n \neq l$, because $\tilde{\lambda}_{1,1}, \tilde{\lambda}_{2,2}, \ldots, \tilde{\lambda}_{N,N}$ are free parameters that can be arbitrarily chosen. Suppose that there exist some m and $l \neq l'$, such that $y_{m,l} \neq 0$ and $y_{m,l'} \neq 0$. Then, $x_{n,m} = 0$ for any $n = 1, 2, \ldots, N$. This contradicts the fact that $\mathbf{O}$ is invertible. Therefore, any row of $\mathbf{O}$ has exactly one non-zero element. Together with the fact that $\mathbf{O}$ is invertible, any column of $\mathbf{O}$ also has exactly one non-zero element. In other words, $\mathbf{O}$ is just a scaling and reordering matrix. By Proposition 4, $\mathrm{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_l^\top \mathbf{R}_0) = 0$ and $\mathrm{cov}(\Delta\tilde{Q}_k^F, \Delta\tilde{Q}_l^F) = 0$ for any $k \neq l$.

To show (43), note that, by (34) and $\mathrm{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_l^\top \mathbf{R}_0) = 0$ for $k \neq l$,

$$\tilde{\mathbf{w}}_k^\top \Delta\mathbf{p}(\Delta\mathbf{Q}) = \sum_{l=1}^{N}\sum_{m=1}^{N} \tilde{\lambda}_{l,m}\Delta\tilde{Q}_m^F \mathrm{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_l^\top \mathbf{R}_0) = \sum_{m=1}^{N} \tilde{\lambda}_{k,m}\Delta\tilde{Q}_m^F \mathrm{var}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0). \tag{B.32}$$

This gives

$$\frac{\partial \tilde{\mathbf{w}}_k^\top \Delta\mathbf{p}(\Delta\mathbf{Q})}{\partial \Delta\tilde{Q}_k^F} = \tilde{\lambda}_{k,k}\mathrm{var}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0). \tag{B.33}$$

Similarly, by (35),

$$\tilde{\mathbf{w}}_k^\top \Delta\bar{\mathbf{p}}(\Delta\mathbf{Q}) = \sum_{m=1}^{N} \tilde{\lambda}_{m,m}\Delta\tilde{Q}_m^F \mathrm{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_m^\top \mathbf{R}_0) = \tilde{\lambda}_{k,k}\Delta\tilde{Q}_k^F \mathrm{var}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0), \tag{B.34}$$

and

$$\frac{\partial \tilde{\mathbf{w}}_k^\top \Delta\bar{\mathbf{p}}(\Delta\mathbf{Q})}{\partial \Delta\tilde{Q}_k^F} = \tilde{\lambda}_{k,k}\mathrm{var}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0). \tag{B.35}$$

This shows that (43) holds.

Under the price impact model (34), the expected compensation for risk is

$$\mathbb{E}[\Delta\mathbf{Q}^\top\Delta\mathbf{p}(\Delta\mathbf{Q})] = \sum_{k=1}^{N}\sum_{l=1}^{N}\tilde{\lambda}_{k,l}\mathrm{var}(\tilde{\mathbf{w}}_k^\top\mathbf{R}_0)\mathrm{cov}(\Delta\tilde{Q}_l^F,\Delta\tilde{Q}_k^F) = \sum_{k=1}^{N}\tilde{\lambda}_{k,k}\mathrm{var}(\tilde{\mathbf{w}}_k^\top\mathbf{R}_0)\mathrm{var}(\Delta\tilde{Q}_k^F), \tag{B.36}$$

where the first equality uses $\mathrm{cov}(\tilde{\mathbf{w}}_k^\top\mathbf{R}_0,\tilde{\mathbf{w}}_l^\top\mathbf{R}_0) = 0$ and $\mathbb{E}[\Delta\mathbf{Q}] = \mathbf{0}$, and the second equality uses $\mathrm{cov}(\Delta\tilde{Q}_k^F,\Delta\tilde{Q}_l^F) = 0$ for any $k \neq l$. Therefore, the expected compensation $\mathbb{E}[\Delta\mathbf{Q}^\top\Delta\mathbf{p}(\Delta\mathbf{Q})]$ under models (34) and (35) are the same for any N^2 parameters $\tilde{\lambda}_{k,l}$. By equation (40), the expected utility $\mathbb{E}[u(W(\Delta\mathbf{Q}))]$ under models (34) and (35) are also the same.

Next, the necessity direction is considered. By (34),

$$\tilde{\mathbf{w}}_k^\top\Delta\mathbf{p}(\Delta\mathbf{Q}) = \sum_{l=1}^{N}\sum_{m=1}^{N}\tilde{\lambda}_{l,m}\Delta\tilde{Q}_m^F\mathrm{cov}(\tilde{\mathbf{w}}_k^\top\mathbf{R}_0,\tilde{\mathbf{w}}_l^\top\mathbf{R}_0). \tag{B.37}$$

Then,

$$\frac{\partial\tilde{\mathbf{w}}_k^\top\Delta\mathbf{p}(\Delta\mathbf{Q})}{\partial\Delta\tilde{Q}_k^F} = \sum_{l=1}^{N}\tilde{\lambda}_{l,k}\mathrm{cov}(\tilde{\mathbf{w}}_k^\top\mathbf{R}_0,\tilde{\mathbf{w}}_l^\top\mathbf{R}_0). \tag{B.38}$$

Similarly, by (35),

$$\tilde{\mathbf{w}}_k^\top\Delta\bar{\mathbf{p}}(\Delta\mathbf{Q}) = \sum_{l=1}^{N}\tilde{\lambda}_{l,l}\Delta\tilde{Q}_l^F\mathrm{cov}(\tilde{\mathbf{w}}_k^\top\mathbf{R}_0,\tilde{\mathbf{w}}_l^\top\mathbf{R}_0), \tag{B.39}$$

and

$$\frac{\partial\tilde{\mathbf{w}}_k^\top\Delta\bar{\mathbf{p}}(\Delta\mathbf{Q})}{\partial\Delta\tilde{Q}_k^F} = \tilde{\lambda}_{k,k}\mathrm{var}(\tilde{\mathbf{w}}_k^\top\mathbf{R}_0). \tag{B.40}$$

Because (43) holds for any parameters $\tilde{\lambda}_{l,k}$, $\mathrm{cov}(\tilde{\mathbf{w}}_k^\top\mathbf{R}_0,\tilde{\mathbf{w}}_l^\top\mathbf{R}_0) = 0$ for any $k \neq l$.

Under the price impact model (34), the expected compensation is

$$\mathbb{E}[\Delta\mathbf{Q}^\top \Delta\mathbf{p}(\Delta\mathbf{Q})] = \sum_{k=1}^{N}\sum_{l=1}^{N} \tilde{\lambda}_{k,l}\mathrm{var}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0)\mathrm{cov}\left(\Delta\tilde{Q}_k^F, \Delta\tilde{Q}_l^F\right), \tag{B.41}$$

which uses $\mathrm{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_l^\top \mathbf{R}_0) = 0$ for any $k \neq l$. By (40), if the expected utility $\mathbb{E}[u(W(\Delta\mathbf{Q}))]$ remains invariant under models (34) and (35), the expected compensation $\mathbb{E}[\Delta\mathbf{Q}^\top \Delta\mathbf{p}(\Delta\mathbf{Q})]$ also remains invariant. Therefore, $\mathrm{cov}(\Delta\tilde{Q}_k^F, \Delta\tilde{Q}_l^F) = 0$ for any $k \neq l$. By Theorem 1, these covariances $\mathrm{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_l^\top \mathbf{R}_0) = 0$ and $\mathrm{cov}(\Delta\tilde{Q}_k^F, \Delta\tilde{Q}_l^F) = 0$ for any $k \neq l$ imply that the orthogonalized model (35) satisfies the IUP for any parameters $\tilde{\lambda}_{1,1}, \tilde{\lambda}_{2,2}, \ldots, \tilde{\lambda}_{N,N}$.

Online Appendix of "An Arbitrage Foundation for Demand Effects in Asset Pricing"

This online appendix contains additional results omitted from the main text.

A General Theory with Duplicate Eigenvalues

This section develops the general theory of the SDF change and price impacts without imposing regularity Assumption 5 in the main text. That is, the matrix $\text{var}(\mathbf{U}\Delta\mathbf{Q})$ may have duplicate eigenvalues.

A.1 Factor Model of SDF Change and Price Impacts

To derive the SDF change in the general case, perform an eigenvalue decomposition:

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

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 $\mathbf{U}\text{var}(\Delta\mathbf{Q})\mathbf{U}^\top$ 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. Pick arbitrary one and construct

the corresponding factors $\mathbf{W} = \mathbf{U}^{-1}\mathbf{G}$ following the proof of Proposition 2 in the main text. Using the N factors $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_N)$ and their corresponding trading demand $\Delta\mathbf{Q}^F = (\Delta Q_1^F, \ldots, \Delta Q_N^F)^\top$, Theorem 1 is extended as follows.

Theorem O.1. *Assumptions 1, 2, 3, and 4 hold if and only if the SDF change can be expressed as*

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

where $\mathbf{\Lambda}$ *is an* $N \times N$ *block-diagonal matrix*

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

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

$$\Delta\mathbf{p}(\Delta\mathbf{Q}) = \mathbb{E}[\mathbf{R}_0 \Delta M(\Delta\mathbf{Q})] = \mathrm{var}(\mathbf{R}_0)\mathbf{W}\mathbf{\Lambda}\Delta\mathbf{Q}^F. \tag{OA.5}$$

Proof. First, verify that the form (OA.3) satisfies the IUP. That is, for any portfolio $\mathbf{a} \in \mathbb{R}^N$, trading demand 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 $\mathrm{cov}(\mathbf{a}^\top \Delta\mathbf{p}(\Delta\mathbf{Q}), s) = 0$.

Project portfolio $\mathbf{a} \in \mathbb{R}^N$ onto the factors $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)$,

$$\mathbf{a} = \sum_{k=1}^{N} x_k \mathbf{w}_k = \mathbf{W}\mathbf{x}, \tag{OA.6}$$

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

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

because of the orthogonalization condition $\mathbf{W}^\top \text{var}(\mathbf{R}_0)\mathbf{W} = \mathbf{I}_N$. Note that $\mathbf{\Lambda}$ is a block-diagonal matrix of the form (OA.4). Therefore, to show $\text{cov}(\mathbf{a}^\top \Delta\mathbf{p}(\Delta\mathbf{Q}), s) = 0$, it is sufficient to show that $x_l \text{cov}(\Delta Q_m^F, s) = 0$ 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 (B.16) in the main text and equation (OA.6), the IUP condition can be simplified as

$$0 = \mathbf{a}^\top \mathbf{C}\mathbf{d} = \mathbf{x}^\top \mathbf{W}^\top \mathbf{C}\mathbf{W}\text{cov}(\Delta\mathbf{Q}^F, s). \tag{OA.8}$$

Define the set

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

Recall that equation (OA.1) defines the matrix $\mathbf{W} = \mathbf{U}^{-1}\mathbf{G}$ from a given of set of eigenvectors $\mathbf{G}$ of matrix $\mathbf{U}\text{var}(\Delta\mathbf{Q})\mathbf{U}^\top$. By the transformation $\mathbf{C} = \mathbf{U}^\top\mathbf{H}\mathbf{U}$, $\mathbf{a}^\top\mathbf{C}\mathbf{d} = 0$ for all $\mathbf{C} \in \mathcal{C}$ if and only if $\mathbf{x}^\top\mathbf{G}^\top\mathbf{H}\mathbf{G}\text{cov}(\Delta\mathbf{Q}^F, s) = 0$ for all $\mathbf{H} \in \mathcal{H}$.

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

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

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

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

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

Moreover, equation (OA.5) implies that

$$\Delta \mathbf{Q}^\top \Delta \mathbf{p}(\Delta \mathbf{Q}) = (\Delta \mathbf{Q}^F)^\top \mathbf{W}^\top \text{var}(\mathbf{R}_0) \mathbf{W} \mathbf{\Lambda} \Delta \mathbf{Q}^F = (\Delta \mathbf{Q}^F)^\top \mathbf{\Lambda} \Delta \mathbf{Q}^F. \tag{OA.12}$$

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

Second, if the SDF change satisfies the IUP, it can be written in the form (OA.3). Following the proof of Theorem 1 in the main text,

$$\Delta M(\Delta \mathbf{Q}) = \sum_{k=1}^{N} \Delta Q_k^F \mathbf{h}_k^\top (\mathbf{R}_0 - \mathbb{E}[\mathbf{R}_0]) \tag{OA.13}$$

for some $\mathbf{h}_k \in \mathbb{R}^N$. The corresponding price impact model is

$$\Delta \mathbf{p}(\Delta \mathbf{Q}) = \text{var}(\mathbf{R}_0) \sum_{k=1}^{N} \Delta Q_k^F \mathbf{h}_k. \tag{OA.14}$$

Apply the IUP with portfolio $\mathbf{a} = \mathbf{w}_l$ and demand $s = \Delta Q_m^F$, 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 (B.16) in the main text,

$$\mathbf{d} = \sum_{k=1}^{N} \mathbf{w}_k \text{cov}(\Delta Q_k^F, \Delta Q_m^F) = \text{var}(\Delta Q_m^F)\mathbf{w}_m. \tag{OA.15}$$

Next, show that $\mathbf{w}_l^\top \mathbf{C}\mathbf{w}_m = 0$ for all $\mathbf{C}$ that satisfies $\text{var}(\mathbf{R}_0)\text{var}(\Delta\mathbf{Q})\mathbf{C} = \mathbf{C}\text{var}(\Delta\mathbf{Q})\text{var}(\mathbf{R}_0)$.

Proof. For any such matrix, define the matrix

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

Then $\mathbf{H}$ satisfies

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

which simplifies to

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

Define the matrix

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

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

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

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

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

Therefore, $\mathbf{Lv}$ is also a vector in the span of the j-th part of the partition. Because one 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, $\mathbf{g}_l^\top \mathbf{H} \mathbf{g}_m = 0$, 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 $\mathbf{W} = \mathbf{U}^{-1}\mathbf{G}$. Therefore, $\mathbf{w}_l^\top \mathbf{C} \mathbf{w}_m = 0$. □

The IUP condition therefore implies that

$$0 = \operatorname{cov}(\mathbf{a}^\top \Delta\mathbf{p}(\Delta\mathbf{Q}), s) = \operatorname{cov}\left(\mathbf{w}_l^\top \operatorname{var}(\mathbf{R}_0) \sum_{k=1}^{N} \Delta Q_k^F \mathbf{h}_k, \Delta Q_m^F \right) = \operatorname{var}(\Delta Q_m^F) \mathbf{w}_l^\top \operatorname{var}(\mathbf{R}_0) \mathbf{h}_m. \tag{OA.23}$$

What remains follows the proof of Theorem 1 in the main text. Specifically, project $\mathbf{h}_m$ onto the factor portfolios $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N$ and obtain $\mathbf{h}_m = \sum_{k=1}^{N} \theta_{m,k} \mathbf{w}_k$. The IUP condition (OA.23) implies that $\theta_{m,k} = 0$, for any (k, k)-th and (m, m)-th elements of matrix $\mathbf{\Pi}$ that correspond to distinct eigenvalues. Therefore, the block-diagonal form of $\mathbf{\Lambda}$ is recovered as in (OA.4). Lastly, because of (OA.12) and Assumption 2, $\mathbf{\Psi}_j$ is positive definite for all j. □

A.2 Optimality of the IUP Orthogonalization

This section establishes the optimality of the IUP orthogonalization, without imposing regularity Assumption 5 in the main text.

In the general case, the definition of model orthogonalization also needs to be modified to account for duplicate eigenvalues. The orthogonalized model now has the block-diagonal form that is consistent with (OA.2).

DEFINITION O.1. *Let* $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_N)$ *be* N *linearly independent portfolios. Expressing the full model in this basis gives*

$$\Delta\mathbf{p}(\Delta\mathbf{Q}) = \mathrm{var}(\mathbf{R}_0)\tilde{\mathbf{W}}\tilde{\mathbf{\Lambda}}\Delta\tilde{\mathbf{Q}}^F, \tag{OA.24}$$

with portfolio-level trading demand $\Delta\tilde{\mathbf{Q}}^F = (\Delta\tilde{Q}_1^F, \Delta\tilde{Q}_2^F, \ldots, \Delta\tilde{Q}_N^F)^\top$ *and the* $N \times N$ *matrix*

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

where each $\tilde{\mathbf{\Psi}}_{j,l}$ *is an* $r_j \times r_l$ *matrix. The orthogonalized* N*-factor model associated with this basis is defined as*

$$\Delta\bar{\mathbf{p}}(\Delta\mathbf{Q}) = \mathrm{var}(\mathbf{R}_0)\tilde{\mathbf{W}}\bar{\mathbf{\Lambda}}\Delta\tilde{\mathbf{Q}}^F, \tag{OA.26}$$

with the $N \times N$ matrix

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

Theorem 2 of the main text is generalized as follows.

Theorem O.2. *Fix any orthogonalization $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots, \tilde{\mathbf{w}}_N)$. Consider the full model (OA.25) with slopes $\tilde{\mathbf{\Lambda}}$ and its orthogonalized counterpart (OA.26) with slopes $\bar{\mathbf{\Lambda}}$. The orthogonalized model satisfies IUP for every choice of $\bar{\mathbf{\Lambda}}$ if and only if the following two conditions hold for every set of $\tilde{\mathbf{\Lambda}}$:*

- *Invariance of own-price sensitivity: for each k,*

$$\frac{\partial \tilde{\mathbf{w}}_k^\top \Delta \mathbf{p}(\Delta \mathbf{Q})}{\partial \Delta \tilde{Q}_k^F} = \frac{\partial \tilde{\mathbf{w}}_k^\top \Delta \bar{\mathbf{p}}(\Delta \mathbf{Q})}{\partial \Delta \tilde{Q}_k^F}. \tag{OA.28}$$

- *The marginal investor's expected utility $\mathbb{E}[u(W(\Delta \mathbf{Q}))]$ is the same under (34) and (35).*

Proof. The proof relies on three intermediate results.

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

Proof. First, the sufficiency direction is established. From Theorem O.1, if the orthogonalized model (OA.26) satisfies the IUP, then there exists some $N \times N$ invertible matrix $\mathbf{O}$, such that for any $\bar{\mathbf{\Lambda}}$ in (OA.27), $\mathbf{O}^{-1}\bar{\mathbf{\Lambda}}\mathbf{O}$ is still a block-diagonal matrix of form (OA.27). Denote

the (j,l)-th block of $\mathbf{O}^{-1}$ as $\mathbf{X}_{j,l}$, which is an $r_j \times r_l$ matrix. Denote the (j,l)-th block of $\mathbf{O}$ as $\mathbf{Y}_{j,l}$, which is an $r_j \times r_l$ matrix. Because $\tilde{\mathbf{\Psi}}_{1,1}, \tilde{\mathbf{\Psi}}_{2,2}, \ldots, \tilde{\mathbf{\Psi}}_{J,J}$ are free parameters that can be arbitrarily chosen, $\mathbf{X}_{j,l}\tilde{\mathbf{\Psi}}_{l,l}\mathbf{Y}_{l,k} = \mathbf{0}$ 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, one can find linearly dependent rows or columns of $\mathbf{O}$, contradicting the invertibility. Therefore, the matrix $\mathbf{O}$ is of the form

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

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

of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues.

Next, the necessity direction is established. Because $\text{cov}(\tilde{\mathbf{w}}_n^\top \mathbf{R}_0, \tilde{\mathbf{w}}_m^\top \mathbf{R}_0) = 0$ and $\text{cov}(\Delta\tilde{Q}_n^F, \Delta\tilde{Q}_m^F) = 0$ whenever the (n,n)-th and (m,m)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues, all one needs for the form (OA.4) is to orthogonalize the portfolios $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots, \tilde{\mathbf{w}}_N)$ within each duplicate eigenvalue. One can always do this because there is no dimension reduction for the $\tilde{\mathbf{\Lambda}}$ matrix within the same eigenvalue. The form (OA.4) satisfies the IUP by Theorem O.1. □

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

Proof. Note that by (OA.24),

$$\tilde{\mathbf{w}}_k^\top \Delta\mathbf{p}(\Delta\mathbf{Q}) = \sum_{l=1}^{N}\sum_{m=1}^{N} \tilde{\lambda}_{l,m} \Delta\tilde{Q}_m^F \text{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_l^\top \mathbf{R}_0). \tag{OA.30}$$

This gives

$$\frac{\partial \tilde{\mathbf{w}}_k^\top \Delta\mathbf{p}(\Delta\mathbf{Q})}{\partial \Delta\tilde{Q}_k^F} = \sum_{l=1}^{N} \tilde{\lambda}_{l,k} \text{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_l^\top \mathbf{R}_0). \tag{OA.31}$$

Similarly, by (OA.26),

$$\tilde{\mathbf{w}}_k^\top \Delta\bar{\mathbf{p}}(\Delta\mathbf{Q}) = \sum_{l=1}^{N}\sum_{m=1}^{N} \bar{\lambda}_{l,m} \Delta\tilde{Q}_m^F \text{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_l^\top \mathbf{R}_0) \tag{OA.32}$$

and

$$\frac{\partial \tilde{\mathbf{w}}_k^\top \Delta\bar{\mathbf{p}}(\Delta\mathbf{Q})}{\partial \Delta\tilde{Q}_k^F} = \sum_{l=1}^{N} \bar{\lambda}_{l,k} \text{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_l^\top \mathbf{R}_0). \tag{OA.33}$$

If the (n,n)-th and (l,l)-th elements of matrix $\mathbf{\Pi}$ correspond to the same eigenvalue, then $\tilde{\lambda}_{n,l} = \bar{\lambda}_{n,l}$ by definition (OA.27). Therefore, (OA.31) and (OA.33) are equal for any $\tilde{\mathbf{\Lambda}}$ if

and only if $\text{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_l^\top \mathbf{R}_0) = 0$ whenever (k,k)-th and (l,l)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues. □

The third claim is stated under the condition that $\text{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_l^\top \mathbf{R}_0) = 0$ whenever the (k,k)-th and (l,l)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues. The expected utility $\mathbb{E}[u(W(\Delta\mathbf{Q}))]$ remains invariant under models (OA.24) and (OA.26) if and only if $\text{cov}(\Delta\tilde{Q}_k^F, \Delta\tilde{Q}_l^F) = 0$ whenever the (k,k)-th and (l,l)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues.

Proof. Partition $\Delta\tilde{\mathbf{Q}}^F$ according to eigenvalue values,

$$\Delta\tilde{\mathbf{Q}}^F = (\Delta\tilde{\mathbf{Q}}_1^F; \Delta\tilde{\mathbf{Q}}_2^F; \ldots; \Delta\tilde{\mathbf{Q}}_J^F), \tag{OA.34}$$

where each $\Delta\tilde{\mathbf{Q}}_j^F$ is an $r_j \times 1$ vector. Similarly, partition $\tilde{\mathbf{W}}^\top \text{var}(\mathbf{R}_0)\tilde{\mathbf{W}}$ as

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

where each $\tilde{\tilde{\mathbf{\Phi}}}_{j,j}$ is an $r_j \times r_j$ matrix and this partition uses $\text{cov}(\tilde{\mathbf{w}}_k^\top \mathbf{R}_0, \tilde{\mathbf{w}}_l^\top \mathbf{R}_0) = 0$ whenever the (k,k)-th and (l,l)-th elements of matrix $\mathbf{\Pi}$ correspond to distinct eigenvalues.

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

$$\mathbb{E}[\Delta\mathbf{Q}^\top \Delta\mathbf{p}(\Delta\mathbf{Q})] = \mathbb{E}[(\Delta\tilde{\mathbf{Q}}^F)^\top \tilde{\mathbf{W}}^\top \text{var}(\mathbf{R}_0)\tilde{\mathbf{W}}\tilde{\mathbf{\Lambda}}\Delta\tilde{\mathbf{Q}}^F] = \sum_{j=1}^{J}\sum_{m=1}^{J} \mathbb{E}\left[(\Delta\tilde{\mathbf{Q}}_j^F)^\top \tilde{\tilde{\mathbf{\Phi}}}_{j,j}\tilde{\tilde{\mathbf{\Psi}}}_{j,m}\Delta\tilde{\mathbf{Q}}_m^F\right]. \tag{OA.36}$$

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

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

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

Together, the three claims establish Theorem O.2. □