Causal Inference for Asset Pricing

Valentin Haddad Zhiguo He Paul Huebner
Peter Kondor Erik Loualiche*

November 2024

Preliminary and Incomplete: Do Not Circulate

Download the latest version here.

Abstract

This paper provides a guide for using causal inference with asset prices and quantities. Our framework revolves around two simple assumptions: homogenous substitution conditional on observables and constant relative elasticity. Under these assumptions, standard cross-sectional instrumental variable or difference-in-difference regressions identify the relative demand elasticity between assets, the difference between own-price and cross-price elasticity. In contrast, identifying aggregate elasticities and substitution along specific characteristics necessarily relies jointly on exogenous sources of time-series variation alone. The same principles also apply to the estimation of multipliers measuring the price impact of supply or demand shocks. The two assumptions map to familiar restrictions on covariance matrices in classical asset pricing models, encompass models from the industrial organization literature such as logit, and accommodate rich substitution patterns even outside of these models. We discuss how to design experiments satisfying these conditions and offer diagnostics to validate them.

*We thank Zhiyu Fu, Ralph Koijen and Jane Li, and seminar participants at Minnesota Carlson, Chicago Booth, UCLA Anderson, Stockholm School of Economics, and Demand in Asset Markets Working Group for helpful comments and suggestions.

Introduction

Causal inference methods that leverage plausibly exogenous sources of variation have become essential tools in empirical economics (Angrist and Pischke, 2009). Recently, these methods have gained traction in asset pricing to better understand the demand for financial assets, through both specific experiments like index inclusions (Shleifer, 1986; Chang et al., 2014) and as building blocks for estimating demand systems (Koijen and Yogo, 2019; Haddad et al., 2024). However, these approaches differ sharply from traditional empirical asset pricing methods, which instead prioritize tests of equilibrium relationships such as Euler equations and factor models (see, e.g., Cochrane, 2005; Campbell, 2017).

We provide a framework for using causal inference in the asset pricing context. Portfolio theory (Markowitz, 1952) teaches us that the demand for various assets is interconnected: the price of all assets affects the demand for all assets. These spillovers are a departure from the canonical causal inference framework, hence additional assumptions are necessary. This paper gives two broad conditions that allow the use of the standard toolbox of causal inference. We completely characterize what sources of variation and estimation procedures identify portfolio demand and its equilibrium impact under these assumptions.

Our goal is to empower researchers in using and interpreting evidence from natural experiments in asset markets as well as understanding their limits. Because our conditions are flexible, they can be used to guide empirical design and simple diagnostics can assess their plausibility in the data, all without having to take a strong stance on a specific model. We also show that the two assumptions map to familiar restrictions on covariance matrices in classical asset pricing models, encompass models from the industrial organization literature such as logit, and accommodate rich substitution patterns even beyond these settings. As such our results also help understand the generality of identification results obtained within structural models (e.g. Koijen and Yogo (2019) or Gabaix and Koijen (2021)) and assess which empirical features matter for addressing specific counterfactual questions.

To understand our framework, start from a naive causal inference approach to estimate how an investor's portfolio decisions respond to prices. An experiment creates an exogenous shift in the price of two groups of assets, treated and controls. The treated assets receives a shock that decreases their price, while the control does not. Here, exogenous means that the shock is unrelated to shifts in the investor's demand curve such as changes in their preferences or their views about the assets. Then, how much does the investor increase their position in a treated asset? To quantify this relation, an elasticity of demand $\widehat{\mathcal{E}}$ is estimated with the following instrumental variable (IV) regression specification of a change in the demand for asset i, ΔD_i in response to a change in its price ΔP_i:

$$\Delta D_i = \widehat{\mathcal{E}} \Delta P_i + \theta' X_i + \epsilon_i, \tag{1}$$

$$\Delta P_i = \lambda Z_i + \eta' X_i + u_i, \tag{2}$$

where X_i are observables, and the instrument Z_i measures the shock to prices and is orthogonal to the residual ϵ_i, conditional on X_i.

From Markowitz (1952), we know that portfolio decisions are not made asset by asset. Financial assets are alternative means of transferring money across states of the world, and thus often close substitutes. Therefore, the prices of all assets affect the demand for all assets

$$\Delta D = \mathcal{E} \Delta P + \epsilon, \tag{3}$$

and $\mathcal{E}$ is an entire matrix of own-price and cross-price elasticities. For example, in the mean-variance setting the elasticity matrix is determined by risk aversion and the covariance matrix of asset returns. In particular, the price of the treated affects the demand for the control, the price of the control affects the demand for the treated, and the prices of every other asset affect both demands. This is the well-known challenge of demand estimation with multiple goods.

Our framework characterizes conditions under which the IV coefficient estimate $\widehat{\mathcal{E}}$ reveals a useful component of the elasticity matrix $\mathcal{E}$ and what this component represents if these conditions are met. When the regression is well specified, it identifies the *relative elasticity* between pairs of assets, the difference between the own-price and cross-price elasticity. For example, in an idealized experiment with perfect randomization — a more stringent condition than standard exogeneity — the coefficient $\widehat{\mathcal{E}}$ is a weighted-average of relative elasticity for all pairs of assets irrespective of the structure of $\mathcal{E}$. The relative elasticity measures how the demand for one asset compared to another one responds to a change in the relative price of these assets. This quantity is particularly valuable for answering micro-level questions hinging on comparison between assets.

We provide two simple conditions to make this insight applicable in practice, that is, when the instrument Z_i only satisfies the standard exogeneity condition of canonical causal inference. Said otherwise, should the researcher find these assumptions appropriate, the discussion of an instrument's validity does not require additional care due to spillovers or equilibrium. The first condition is *constant relative elasticity*: the difference between own and cross-price elasticity for two assets with the same observable characteristics is constant in the estimation sample. This condition ensures that there is a single number to estimate.[1] The second condition is *homogenous substitution*: the demand for all assets in the estimation with the same characteristics must react similarly to the price of all other assets in the investor's investment set. For example, your demand for Ford and General Motors respond in the same way to the price of Netflix. Importantly, this condition must apply both to substitution with respect to other assets inside the estimation and outside of the sample.[2] Without such an assumption, variations in the price of these other assets can create heterogenous spillovers which act as omitted variables in the regression. If those two conditions are verified, jointly with the usual exogeneity and relevance conditions, the IV coefficient $\widehat{\mathcal{E}}$ of equation (1) estimates the relative elasticity in the sample.

[1] We favor this simple setting as opposed to tracking an average effect because standard regression methods do not lead to average treatment effect estimation in presence of controls. Goldsmith-Pinkham et al. (2022) explain this challenge and proposes some alternative estimation approaches.

[2] Excluded assets outside of the sample can arise because of the common issue that the econometrician does not observe all of the investor's holdings, or by design so as to make the two conditions more plausible in the sample.

Despite their simple statement, these conditions accommodate a large variety of portfolio demands, that is, shapes of the elasticity matrix $\mathcal{E}$. This versatility allows the framework to fit different types of natural experiments.

In the mean-variance framework, applying these two conditions without observable characteristics has straightforward implications on asset return covariance matrix. Constant relative elasticity and homogenous substitution correspond to the assumption that treated and controls have similar variances and covariances. Moreover, homogenous substitution implies that all assets in the sample share the same covariance with each of the excluded assets. While these conditions do not hold for arbitrary sets of assets, they can guide the researcher to design an appropriate estimation sample. For example, the researcher could consider a small group of assets with similar characteristics: in the same industry, with the same size, etc. Then, they could check if these assets indeed have similar volatility and covariance with each other. While it is not possible to assess the covariance with every possible excluded asset, they could present evidence of similar covariances with various portfolios of assets, that is evidence of balance in betas.

This basic approach is particularly suitable for settings with local natural experiments. Controlling for observables open up the possibility of larger samples with more heterogeneity. One could have multiple groups of assets, where elasticities between assets in each group are symmetric, but elasticities across groups differ. In this case, one can still estimate the relative elasticity within group by including group fixed effects. Heterogeneity in substitution can also be dealt with if its determinants are known. For example, in the mean-variance setting, heterogenous betas on common factors create such a phenomenon, and these betas must be controlled for. Heterogenous substitution can also be driven by other motives than risk. An investor might balance their portfolio's carbon emissions or target some regulatory constraint, and hence substitute across assets based on the corresponding characteristic for each asset.

The identification result also helps understand the generality of properties previously derived under specific models. In a seminal contribution, Koijen and Yogo (2019) show that a similar estimation equation arises if one either assumes that investors have mean-variance investors with beliefs on factor loadings and expected returns that are functions of characteristic, or that they have logit demand. This occurs because both models satisfy the two conditions of this paper. Moreover, our framework shows that the interpretation of the IV coefficient as a relative elasticity remains valid for many models outside of these two.

Another natural application of causal inference is the measurement of multipliers or price impacts: the effect of an exogenous supply or demand shock for an asset on its price. This type of question flips prices and quantities relative to demand estimation. For example, how do Fed asset purchases affect Treasury prices? A similar issue as for demand elasticities arises naturally: there is no such thing as the multiplier but instead a multiplier matrix whereby the demand for all assets affect the price of all assets. The two conditions we put forward for demand estimation also apply to this context, and under these conditions, causal inference measures a relative multiplier: the effect of changing the supply of one asset relative to another on the price of this asset relative to the other. The connection between elasticity and multiplier is stronger: in equilibrium, the multiplier matrix $\mathcal{M}$ is the inverse of the aggregate demand elasticity matrix. Neither individual own-price nor cross-price elasticities and multiplier are the inverse of each other under matrix inversion. However, we show that under the two assumptions, relative elasticity and multiplier have such a relation, $\widehat{\mathcal{M}} = \widehat{\mathcal{E}}^{-1}$, and therefore both estimation approaches convey the same information.

The relative elasticity is only one coefficient — or more precisely a single combination of the matrix coefficients — and hence it is not the full matrix. In other words, the cross-sectional regression alone is not enough to answer questions beyond local comparisons between assets. The cross-section is not enough to separate own- and cross-price elasticity even if they are constant. The cross-section also fails to reveal substitution across broad categories: how do you rebalance when the price of small stocks relative to big stocks changes, or when the price of long-duration with respect to short-duration bonds changes. Finally, the cross-section is not sufficient to characterize the price impact of a demand shock for all assets at the same time, the classic problem of "missing intercept" in going from micro to macro estimates.

A natural starting point to answer this type of meso- or macro-level questions is to use data aggregated across assets. For example, instead of keeping track of demand and price stock by stock, researchers focus on a relation between the overall demand for stocks and a price index for stocks. Gabaix and Koijen (2021) present such a framework and estimate the macro multiplier, that is, the price impact of a demand shock for all stocks. Because there is only one asset (the market index), one must rely on exogenous source of variation in the time series.

Starting from an asset-level elasticity matrix that satisfies our two conditions, we ask when it is justified to focus on the aggregated data. First, we show that when substitution is homogenous unconditionally, that is, when cross-elasticities are constant, demand can be decomposed into an aggregate component and an asset-level component. In the aggregate component, the overall demand for assets responds to a price index for all assets. In the asset-level component, the demand for each asset relative to the index only depends on its price relative to the index. This decomposition leads to a separation in identification: the relative elasticity is estimated from the cross-section alone, while the macro elasticity is estimated from the time series alone. Interestingly, this result also highlights that separating own and cross-price elasticity must also rely on the time series.

When substitution is richer and depends on observables, one must track aggregates of price and demand along multiple dimensions. These aggregates are a counterpart to factors in standard asset pricing, with the demand for factors responding to their price of risk. For example, if substitution across bonds depends on their duration, the duration-weighted portfolio matters beyond the market portfolio. Demand can then be separated between an asset-level component, to be estimated from the cross-section, and a component depending on all of the aggregates estimated, to be estimated from the time series. However, how the demand for the various aggregates respond to their price cannot be separated from each other in general.[3]

To understand this interconnection, go back to the example of bonds. Consider a "twist" demand shock that buys long-term bonds and sells short-term bond with zero net purchase; a standard quantitative easing operation. When duration matters for substitution, such a shock can affect the price of the market index for bonds despite creating no shift in the overall demand for bonds. If such twist shocks are correlated with shocks in the aggregate demand for bonds, they would act as an omitted variable in a regression of the market index on aggregate shifts in demands, which would lead to biased estimates of the macro multiplier. To alleviate this concern, the econometrician could inspect the composition of their instrument for aggregate demand shocks at the bond level and check that it occurs in a parallel way across maturities. Alternatively, they could include another instrument to separate the impact of aggregate demand shock and twist shocks.

Another important implication of these aggregation results is that it is not enough to estimate a relative (or micro) and a macro elasticity to answer relative questions across broad groups that affect substitutions. Instead, one must directly tackle the elasticity or multiplier along this exact dimension. For instance, to measure the price impact of a shift in demand from brown stocks to green stocks, one must understand how investors substitute across these categories, which can only be revealed using exogenous time series variation in demand for a green-weighted index. Similarly, this concern highlights that empirical demand models should not only include how demand respond to characteristics, but also how elastic it is to the market price of the characteristic.

Taken together, our results offer a user guide for causal inference in asset pricing. We spell out: a) which assumptions one needs to defend to use these estimation techniques, b) diagnostics to assess the plausibility of these assumptions, c) which technique and source of variation is appropriate for different economic questions, d) how to interpret causal estimates.

[3] Huber (2023) highlights a related point in the context of general equilibrium spillovers of large-scale shocks.

Related Literature. *To be completed.*

1 Causal Inference versus Asset Pricing

We set up the basic regression framework for estimating the demand for assets using causal inference. We contrast this setting with how standard asset pricing theory works. The key distinction is the emphasis on strong patterns of substitution across assets.

1.1 The causal inference framework

We focus on a generic setting of identifying the demand for financial assets. Section 3 considers the related problem of price impact from demand shocks. Intuitively, we want to understand how an investor's demand for an asset responds to the price of this asset. We consider the following experiment: a shock exogenous to demand happens and affects the price P_i of various assets indexed by i, with intensity Z_i.

Inspired by standard causal inference, running an instrumental variable estimation on a sample $\mathcal{S}$ of assets is natural in this setting. In this model, one regresses the change in demand for each asset ΔD_i on the change in the price of this asset ΔP_i, using Z_i as an instrument for the change in price. This corresponds to the two-stage least square specification:

$$\Delta D_i = \widehat{\mathcal{E}} \Delta P_i + \theta' X_i + \epsilon_i, \tag{4}$$

$$\Delta P_i = \lambda Z_i + \eta' X_i + u_i, \tag{5}$$

where X is a set of controls to be specified.

The two standard conditions for this regression model to be identified are the relevance and exclusion restrictions. Relevance is the idea that the instrument Z_i creates variation in prices: $\lambda \neq 0$. Exclusion is the idea that the instrument does not affect demand through other channels than the price: $Z_i \perp \epsilon_i | X_i$.

One can imagine running this specification in levels or in logs depending on the model of demand. For example, models like CARA preferences are better behaved in levels, while logit demand matches with logs. We abuse the language of demand estimation slightly and call coefficients in such regressions demand elasticities irrespective of log or levels. Section 2.4 reviews the appropriate units for various standard models. Also while we focus on writing things in changes to match the standard difference-in-difference framework, similar arguments apply without changes.

A simpler benchmark To better understand the behavior of this regression, it is useful to study a simplified version. There is no shift in demand curve, but simply a shock that triggered movements in prices, with the movement in the price of asset 1 larger than in the price of asset 2. There are still many other assets $(3, \ldots, N)$ that might also experience changes in price. For example, the shock could be a surprise increase in the supply of asset 1 but not asset 2. In this case the counterpart to the IV estimator is the relative change in demand for assets 1 and 2 divided by the relative change in price:

$$\mathcal{E} = \frac{\Delta D_1 - \Delta D_2}{\Delta P_1 - \Delta P_2}. \tag{6}$$

To see this result, note that the sample is just the two assets $\mathcal{S} = \{1, 2\}$, the instrument representing the experiment is $Z_1 = 1$ and $Z_2 = 0$ and that there are no controls or error terms.[4]

1.2 Standard asset pricing structure

The setting of equation (4) differs sharply from how standard asset pricing theory specifies the demand for assets. A key insight going back to Markowitz (1952) is that assets are not distinct goods, but instead alternative means of saving with different risk and reward.

[4]The first stage regresses ΔP_i on the dummy, so $\lambda = \Delta P_1 - \Delta P_2$. The second stage regresses the change in demand on the dummy scaled by λ. Without scaling the coefficient would be $\Delta D_1 - \Delta D_2$. With scaling, we obtain equation (6).

Investors choose portfolios optimally combining these assets. This substitutability implies that the demand for one asset depends not only on its own price but also the price of other assets. How many shares of Apple you purchase depends on the price of Apple and also on the price of Nvidia.

The most standard example of this approach is mean-variance optimization: an investor chooses their portfolio to maximize $\mathbb{E}(W) - \frac{\gamma}{2}\operatorname{var}(W)$ where W is their future wealth, and γ measures their risk aversion. If assets have mean payoffs M and covariance matrix $\mathit{\Sigma}$, the vector of demand is:

$$D = \frac{1}{\gamma}\mathit{\Sigma}^{-1}(M - P). \tag{7}$$

Absent demand shocks, this implies that changes in demand can be written as

$$\mathit{\Delta} D = \mathcal{E}\mathit{\Delta} P \iff \mathit{\Delta} D_i = \sum_j \mathcal{E}_{ij}\mathit{\Delta} P_j \tag{8}$$

with the matrix of elasticity $\mathcal{E}$ being determined by risk aversion and the covariance between assets: $\mathcal{E} = -\mathit{\Sigma}^{-1}/\gamma$. When assets are correlated with each other, they become close substitutes and their demand respond to each other's price.

More generally, any model of asset demand will imply its matrix of elasticities $\mathcal{E}$. The diagonal elements of $\mathcal{E}$ measure the own-price elasticities, while the off-diagonal capture cross-price elasticities. If the model is not linear (or log-linear) in prices, we focus on a local approximation of demand.[5]

1.3 The challenge

The distinction between the two approaches is clearly visible: causal inference focuses on a univariate relation between price and demand — the coefficient $\widehat{\mathcal{E}}$ — while standard asset

[5]Such a completely flexible elasticity matrix is reminiscent of the almost-ideal demand system of Deaton and Muellbauer (1980).

pricing emphasizes a multivariate relation — the matrix $\mathcal{E}$. This univariate focus is a key element of standard causal inference: under the stable unit treatment value assumption (SUTVA), treatment on one unit (for us, an asset) does not affect other units.

Concretely, the presence of cross-elasticities imply that the prices of all other assets are omitted variables in equation (4). When we have non-zero elasticity of substitutions between assets, the change in the price of the other assets affect the demand for the original asset. In changes, the demand system of equation (8) gives:

$$\Delta D_i = \mathcal{E}_{ii} \Delta P_i + \sum_{j \neq i} \mathcal{E}_{ij} \Delta P_j + \epsilon_i. \tag{9}$$

A standard natural experiment focuses on a situation where the instrument is orthogonal to shifts in demand, so $Z_i \perp \epsilon_i$. However, other prices naturally respond to the treatment so the other terms in the sum create an omitted variable bias. Fuchs et al. (2024) discusses in length the theoretical foundations of this challenge.

As an example, go back to the deterministic case comparing two assets 1 and 2, and consider the effect of the change in the price of a third asset, say asset 3. This change results in a contribution $(\mathcal{E}_{13} - \mathcal{E}_{23}) \Delta P_3$ to the numerator of equation (6). If the two cross-elasticities differ from each other, this leads to a bias away from the own elasticity. This is the standard problem of demand estimation with multiple goods.

In the face of this challenge one can deem causal inference hopeless for asset pricing and throw their hands in the air. However there is a more constructive approach: acknowledge that additional assumptions about the nature of spillovers are necessary, and that the coefficient $\widehat{\mathcal{E}}$ will only reveal a specific dimension of the matrix $\mathcal{E}$. After all, this is the second part of Markowitz' argument: basic economics can inform us about the structure of substitution across assets. In the rest of the paper, we put forward simple flexible conditions guided by these economic principles. An alternative, more along the lines of modern empirical industrial organization literature, would be to fully specify a structural model. We show later on how our results intersect with this approach.

2 Making Causal Inference Work with Asset Pricing

We provide a framework for using the cross-sectional causal inference regressions in asset pricing We give two natural conditions on the structure of substitution that are sufficient for these regressions to identify a meaningful quantity. These conditions have a simple interpretation in terms of the statistical structure of asset returns in the context of risk-based models. They also accomodate alternative approaches.

2.1 Conditions for valid estimation

We state the two conditions leading to valid estimation. First, we put some structure on substitution between assets in the sample with respect to assets both inside and outside the sample, conditional on observables.

Assumption A1 (Homogenous substitution between assets) *Any pair of assets in the estimation sample with the same observables shares the same cross-price elasticity with respect to each third asset, within or outside of the estimation sample:*

$$X_i = X_j \Rightarrow \mathcal{E}_{il} = \mathcal{E}_{jl} = \mathcal{E}_{cross}(X_i, X_l) = X_i' \mathcal{E}_S X_l, \quad \forall i, j \in \mathcal{S}, l \neq i, j, \tag{10}$$

where X_i is the $K \times 1$ vector of observables for asset i, and $\mathcal{E}_S$ is a $K \times K$ matrix that determines substitution between assets based on their observables vectors.

Since based on condition A1 the cross-price elasticity is a function of the observables, we write it as $\mathcal{E}_{\text{cross}}(X_i, X_l)$. We furthermore parametrize this function as a bilinear form in observables, $\mathcal{E}_{\text{cross}}(X_i, X_l) = X_i' \mathcal{E}_S X_l$, encompassing a large space of potential substitution patterns. If one further assumes that the substitution matrix $\mathcal{E}_S$ is symmetric, then the entire elasticity matrix $\mathcal{E}$ is symmetric as well, consistent with mean-variance models of investor demand, in which substitution works through the inverse covariance matrix.

Assumption A1 states that for two assets that are comparable along every observable, if the price of any third asset, either within or outside the estimation sample, moves, then substitution between the third asset and the two comparable assets is the same. That is, for the pair of comparable companies Ford and General Motors, if the price of any third asset moves, substitution between the third asset and Ford will be same as substitution between the third asset and General Motors. This assumption deals with the omitted variable problem coming from the substitution effect due to the change in other asset prices. It makes this substitution response identical for all assets in the sample when conditioning on the vector of observables X_i, hence absorbed into regression coefficients on the observables in a cross-sectional regression.

The second assumption ensures there is a single number to estimate.

Assumption A2 (Constant relative elasticity) *Assets in the estimation sample have the same relative elasticities:*

$$\mathcal{E}_{ii} - \mathcal{E}_{cross}(X_i, X_i) = \mathcal{E}_{jj} - \mathcal{E}_{cross}(X_j, X_j), \quad \forall i, j \in \mathcal{S} \tag{11}$$

It is natural to assume that something stays fixed across the sample to be able to run a regression.[6] What is more surprising here, is that the restriction applies not to the own-price elasticity but more broadly to the difference of own-price and cross-price elasticity with a comparable asset with the same observables.

Both conditions can be viewed as guidance for the econometrician to choose their sample and their observables appropriately. For example, if they consider a large set of stocks, there will be a lot of heterogeneity in risk, and how the stocks comove with one another. To the extent that the econometrician can capture this heterogeneity in risk through observables,

[6] In the language of causal inference methods, this amounts to a homogeneous treatment effect. In section 2.3.4, we consider heterogeneous treatment effects, in which case the IV estimator estimates a weighted average of relative elasticities.

for example factor exposures or stock characteristics, they should make sure to include such observables. If stocks additionally vary along unobservable dimensions, a focus on a subsample of assets that are naturally more comparable with each other and that arguably does not vary along unobservable dimensions provides an alternative path. We expound these considerations in Section 2.3.

2.2 What does it estimate?

We are now ready to state our main proposition.

Proposition 1 *Under assumptions A1 and A2, as well as the standard relevance and exclustion restrictions, the two-stage least square estimation of equations* (4) *and* (5) *identifies the relative elasticity:*

$$\widehat{\mathcal{E}} = \mathcal{E}_{ii} - \mathcal{E}_{cross}(X_i, X_i), \ \forall i \in \mathcal{S} \tag{12}$$

When the IV estimation is well specified, it identifies the difference between the own-price elasticity and the cross-price elasticity for two assets in the sample with the same observables. While this result stands in contrast from the intuition of measuring "how demand for each asset responds to its own price," it is natural. A cross-sectional regression is a comparison across assets in the sample. Even if only the price of one asset is shocked, the regression coefficient will be driven by the relative response of demand for this asset to demand for the comparable control asset, hence the relative intensity of the own- and cross-elasticity conditional on observables. In other words, $\widehat{\mathcal{E}}$ answers the question: how does the demand for one asset relative to another — comparable — asset respond to the relative price of these assets?

Intuitively, if one knows the variables driving heterogeneity in substitution, the observables X_i, these variables need to be included as controls in the regression, to absorb the variation across assets driven by the common factors, an observation present in Koijen and

Yogo (2019).[7] Once we control for X_i, the regression is equivalent to making pairwise comparisons of assets that have the same observables. As we will show next, this means that substitutions with any third asset are differenced out, effectively solving the challenge created by rich substitution patterns and equilibrium spillovers.

Importantly, there is no sense in which the coefficient θ on the observables X_i identifies anything about the substitution matrix $\mathcal{E}_S$. This is because one cannot separately identify common dimensions of substitution absorbed into θ from outright demand for characteristics and changes in price along these dimensions of substitution. Therefore, it is not possible to make inference about the role of observables X_i for the elasticity matrix using this regression. Section 4 shows how to identify these other dimensions.

Proof for the simple case. To understand the mechanics of this result, let us go back to the deterministic case comparing 2 assets, but imagine that both share the same vector of observables, $X_1 = X_2$. For example, think again about Ford and General Motors. The changes in demands are:

$$\Delta D_1 = \mathcal{E}_{11}\Delta P_1 + \mathcal{E}_{12}\Delta P_2 + \sum_{k>2} \mathcal{E}_{1k}\Delta P_k \tag{13}$$

$$\Delta D_2 = \mathcal{E}_{22}\Delta P_2 + \mathcal{E}_{21}\Delta P_1 + \sum_{k>2} \mathcal{E}_{2k}\Delta P_k \tag{14}$$

A1 implies that the cross-elasticities can be written as functions of observables, and using $X_1 = X_2$, we have:

$$\Delta D_1 = \mathcal{E}_{11}\Delta P_1 + \mathcal{E}_{\text{cross}}(X_1, X_2)\Delta P_2 + \sum_{k>2} \mathcal{E}_{\text{cross}}(X_1, X_k)\Delta P_k \tag{15}$$

$$= \mathcal{E}_{11}\Delta P_1 + \mathcal{E}_{\text{cross}}(X_2, X_2)\Delta P_2 + \sum_{k>2} \mathcal{E}_{\text{cross}}(X_1, X_k)\Delta P_k \tag{16}$$

$$\Delta D_2 = \mathcal{E}_{22}\Delta P_2 + \mathcal{E}_{\text{cross}}(X_1, X_1)\Delta P_1 + \sum_{k>2} \mathcal{E}_{\text{cross}}(X_1, X_k)\Delta P_k \tag{17}$$

[7] Koijen and Yogo (2019) recommend using stock characteristics as proxies for the factor exposures.

Since both assets have the same observable vectors, the subsitution between any third asset and the two of them is identical, and differences out. The difference $\varDelta D_1 - \varDelta D_2$ is equal to:

$$\varDelta D_1 - \varDelta D_2 = \underbrace{(\mathcal{E}_{11} - \mathcal{E}_{\text{cross}}(X_1, X_1))}_{\widehat{\mathcal{E}}} \varDelta P_1 - \underbrace{(\mathcal{E}_{22} - \mathcal{E}_{\text{cross}}(X_2, X_2))}_{\widehat{\mathcal{E}}} \varDelta P_2 \tag{18}$$

$$= \widehat{\mathcal{E}} \left(\varDelta P_1 - \varDelta P_2 \right) \tag{19}$$

The last step uses A2 of identical relative elasticities, $\widehat{\mathcal{E}}$, for both assets 1 and 2. Both the response of the demand to the own price (controlled by $\mathcal{E}_{11}$) and the response to the price of the other asset asset (controlled by $\mathcal{E}_{\text{cross}}(X_1, X_2)$) shape this comparison. However, the relative shock $(\varDelta P_1 - \varDelta P_2)$ hits these mechanisms in the opposite direction. Hence, the regression coefficient is the difference between the own-price elasticity and the cross-price elasticity with a comparable control. Appendix A.2 presents the full proof of the proposition.

2.3 Using the identification results

Assumptions A1 and A2 provide general conditions for the identification with observable-based susbtitution. To use this result, the econometrician must take a stand on the estimation sample and the relevant observables. We discuss a few different approaches to do so, with choices that are both intuitive practically and frequently arise in finance theory.

2.3.1 Homogeneous estimation sample

Consider an econometrician who wants to assess the effects of a local experiment among a sample of few, but highly comparable assets. If the assets in the estimation sample are comparable along every observable, the only control needed for Proposition 1 to hold is a cross-sectional constant. For example, firms in a narrowly defined industry might have similar risk and similar relation with stocks in other industries. Another example could be multiple corporate bonds from the same issuer with similar maturity (see Coppola (2021)).

In this case, assumption A1 implies that the cross-price elasticity is the same for all assets in the estimation sample, while assumption A2 additionally implies that the own-price elasticity is the same for all assets in the estimation sample:

$$\mathcal{E}_{ii} = \mathcal{E}_{own}, \ \forall \ i \in \mathcal{S}, \tag{20}$$

$$\mathcal{E}_{ij} = \mathcal{E}_{cross}, \ \forall i, j \in \mathcal{S}. \tag{21}$$

Naturally, substitution between the assets in the estimation sample and the — in the case of a narrowly defined estimation sample — many outside assets, will not generally be constant. But, it will be the same across assets in the estimation sample with any given outside asset. Technically, observables are unrestricted for the assets outside the estimation sample, but constant within the estimation sample, which together implies that each asset within the estimation sample has the same cross-elasticity with any given asset outside the estimation sample. Figure 1 illustrates such an elasticity matrix.

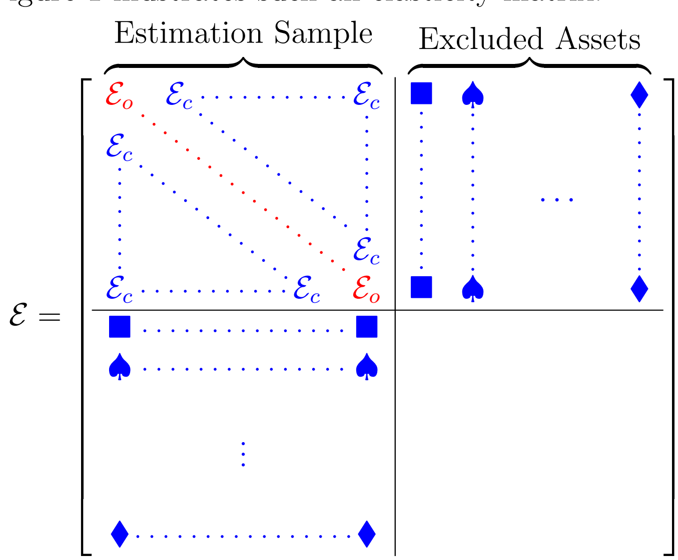

Figure 1: Assumptions A1 (blue) and A2 (red) for an elasticity matrix for local experiments.

In risk-based models, in which elasticities are proportional to the inverse of the covariance matrix, this means that all assets in the sample have the same variance and covariance with each-other. It also corresponds to assuming that for any outside asset k, the covariance of k with any assets in the sample is constant: $\text{cov}(R_i, R_k) = \text{cov}(R_j, R_k)$. Since typically, there are many outside assets, this condition is hard to assess empirically. Still, while making this assumption it is useful to present some corroborating evidence: for example, covariances with a set of known factors.

2.3.2 Groups of assets

One might be able to find not one but multiple groups of assets individually homogeneous, meaning that within-group, substitution looks like in 2.3.1, but across-groups, it does not. For example, this might hold for a set of firms in a narrow industry but not across these industries. In this case, it is possible to get estimates by pooling all industries while including group fixed effects to focus on within-group variation. Within the general setting of Assumption A1, this is the special case where the observables X_i are group dummies. The two-stage least squares regressions takes the form:

$$\mathit{\Delta} D_i = \mathcal{E}\mathit{\Delta} P_i + \theta_g + \epsilon_i, \tag{22}$$

$$\mathit{\Delta} P_i = \lambda Z_i + \eta_g + u_i, \tag{23}$$

where g is the group of assets which includes i, and θ_g, η_g are group fixed effects. Since this is a special case of Assumption A1, the general identification result of Proposition 1 holds when including group-fixed effects.

Chaudhary et al. (2022) explain how not including group-fixed effects in such a situation leads to biased inference. They document the relevance of this bias when measuring the effect of fund flows on corporate bond prices.

2.3.3 Factor models

Some sources of heterogeneity cut across assets. For example, in a downturn, when many asset prices fall, the demand for a more cyclical asset might change differently from that of a less cyclical asset. Empirical asset pricing often highlights many such sources of heterogeneity with risk factors. Even if we focus on a specific industry, different assets might have a different response to inflation or duration. These factors naturally affect patterns of substitutability, because they affect the covariance matrix of returns.

A factor representation of the entire substitution matrix looks very similar to the assumptions, but has a specific interpretation in the context of risk-based asset pricing models:

$$\mathcal{E} = \widehat{\mathcal{E}} I + \beta \Psi \beta', \tag{24}$$

where $\beta = \left[1, \beta^{(f_1)}, \beta^{(f_2)}, \ldots, \beta^{(f_{K-1})}\right]$ is the set of factor loadings, and Ψ a $K \times K$ symmetric matrix. Again, this is a version of the general assumptions A1 and A2, where the $N \times K$ observable matrix X is the factor loadings β, and the substitution matrix $\mathcal{E}_S$ is Ψ. So again, Proposition 1 applies; it is enough to control for β to identify the relative elasticity $\widehat{\mathcal{E}}$.

In the mean-variance framework, this case is equivalent to assuming that there is a set of common factors F_t and with loadings β:

$$R_{i,t} = \beta_i' F_t + \epsilon_{i,t}, \epsilon_i \perp \epsilon_j, \text{ for } i \neq j. \tag{25}$$

The corresponding covariance matrix is $\Sigma = \sigma_\epsilon^2 I + \beta \Sigma_F \beta'$, and similarly for the elasticity matrix $\mathcal{E} = \gamma^{-1} \Sigma^{-1}$.

A variation of this result particularly well-suited for event-study settings is to construct synthetic controls. If one has a set of treated assets, they can construct portfolios of other assets as the control group for a difference-in-difference study. There are two requirements for this approach to be valid. First, the factor exposures of the control portfolio must be the same as that of the treated asset. Second, each asset in the control portfolio (as opposed to the combined portfolio returns) must have the same residual volatility as the treated asset.

Outside of risk-based factor models, investor might be trying to balance other characteristics of the assets in their portfolio. Institutions might specialize in certain asset classes or industries, and have mandates to do so. Or, some investors might have preferences for non-pecuniary aspects of stocks, such as their ESG characteristics. Standard finance theory offers fewer direct guidelines on empirical specifications of such demand.

2.3.4 Unobserved Heterogeneity.

Finally, at the cost of stronger assumptions, one can entertain unobserved heterogeneity in the elasticity matrix $\mathcal{E}$.

Assumption A3 (Unobserved heterogeneity)

A3.a *Data generating process of the first stage follows*

$$\Delta P_i = \lambda_i Z_i + u_i, \quad \textit{with } Z_i \textit{ independent of } (u_i, \lambda_i). \tag{26}$$

A3.b *Homogeneity of the elasticity with respect to the instrument:*

$$(\mathcal{E}_{ii}, \mathcal{E}_{ij})|Z_i \sim (\mathcal{E}_{ii}, \mathcal{E}_{ij}), \ \forall i \in \mathcal{S} \tag{27}$$

Assumption A3 entertains variation in how the instrument transmits to prices (λ_i), and in elasticity across assets. However it imposes that this variation is independent from the instrument itself.

Assumption A4 (Homogeneous substitution from excluded alternatives) *Assets in the sample share the same cross-price elasticity with respect to each asset outside of the estimation sample.*

$$\mathcal{E}_{ik} = \mathcal{E}_{jk}, \ \forall \ i, j \in \mathcal{S}, \ k \notin \mathcal{S} \tag{28}$$

Assumption A4 deals with the omitted variable problem coming from the substitution effect due to the change in unobserved asset prices. It makes this substitution response identical for all assets in the sample, hence a constant absorbed in a cross-sectional regression.

A situation under which considering this type of substitution is appealing is when one believes there is an amount of noise around Assumption A2. Naturally, all assets even in a narrowly-defined group as in 2.3.1 are not exactly identical and small variation in elasticities that one has no reason to believe relates to the experiment at hand fits Assumption A3. Another case where this proposition is useful is when the experiment uses pairs of assets that are different on many dimensions, but in a plausibly random way. An example of this case are index inclusions: the included and excluded assets from the index are closely related to their following stock in size, but might be in different industry, or have different characteristics.

The following proposition shows that assumptions A3 and A4 lead to estimating an average value of relative elasticity.

Proposition 2 *Under the unobserved heterogeneity assumptions A3 and A4, the two-stage least square estimation of equations* (4) *and* (5) *without observables identifies the local average of the relative elasticity:*

$$\widehat{\mathcal{E}} = \frac{\mathbf{E}_i \left\{ \lambda_i (\mathcal{E}_{ii} - \mathbf{E}_j(\mathcal{E}_{ji})) \right\}}{\mathbf{E}_i(\lambda_i)}. \tag{29}$$

With an added monotonicity condition that the instrument always affect prices in the same direction — λ_i of constant sign — the estimate $\widehat{\mathcal{E}}$ will fall within the range of estimates in the sample. If one expects little variation in elasticity this result indicates that heterogeneity will not create large deviation from a situation with exactly constant elasticity. With a wider range of variation in relative elasticity, it becomes interesting to inspect weighting in the average formula. We see that assets for which the instrument have a larger impact on prices (large λ_i) see their relative elasticity receive a higher weight. For example, if more illiquid assets have both a higher impact of the instrument λ_i and investor trade them more inelastically (lower $\mathcal{E}_{ii}$), estimates of relative elasticity will be lower than the unweighted average relative elasticity, and overstate how inelastic the typical asset is.

2.4 Elasticity in standard models of finance

We discuss how standard model of asset demands relate with the identification assumptions. For each model, we derive the appropriate units under which the demand regression is well specified.

In the mean-variance model (CARA) described above, we have seen the direct mapping between covariance matrix and the elasticity matrix when considering a relation between the level of demand the level of prices: $\mathcal{E} = \gamma^{-1}\Sigma^{-1}$.

We turn to the case of a constant relative risk aversion (CRRA) model and the logit model. Table 1 summarizes the results for the three models.

	CARA	CRRA	Logit
Regression units "demand" LHS	demand D_i	portfolio shares P_iD_i/I	log portfolio shares $\log(P_iD_i/I)$
Regression units "price" RHS	price P_i	log price $\log P_i$	log price $\log P_i$
Own Price Elasticity $\mathcal{E}_{ii}$	$\frac{R_f}{\gamma}\Sigma^{-1}_{ii}$	$\frac{1}{\gamma}\Sigma^{-1}_{ii}$	$\alpha(1-\omega_i)$
Cross price Elasticity $\mathcal{E}_{ij}$	$\frac{R_f}{\gamma}\Sigma^{-1}_{ij}$	$\frac{1}{\gamma}\Sigma^{-1}_{ij}$	$-\alpha\omega_j$
Relative Elasticity $\hat{\mathcal{E}} = \mathcal{E}_{ii} - \mathcal{E}_{ji}$	$\frac{R_f}{\gamma}\left(\Sigma^{-1}_{ii} - \Sigma^{-1}_{ji}\right)$	$\frac{1}{\gamma}\left(\Sigma^{-1}_{ii} - \Sigma^{-1}_{ji}\right)$	α
Difference Marshallian vs. Hicksian	No	No	No

Table 1: Three standard cases of demand models.

2.4.1 CRRA

Utility in the Constant Relative Risk Aversion case is given by $u(C) = C^{1-\gamma}/(1-\gamma)$. The risk-free rate has price $P_0 = 1/R_f$ and pays off 1; there are N assets with payoffs $X = \{X_i\}_i$ at time 1, with prices $\{P_i\}$. Returns are $R_i = X_i/P_i$.

To solve for the optimal demands, we assume that the payoffs follow a lognormal distribution: $\log X \sim \mathcal{N}(\mu, \Sigma)$, which gives us lognormal returns as:

$$\log R = \log X - \log P \sim \mathcal{N}(\mu - \log P, \Sigma) \tag{30}$$

Further we log-linearize portfolio returns following Campbell and Viceira (2002).[8]

For investor i with wealth W, the optimal demand is:

$$D_i = \frac{1}{\gamma}\frac{W}{P_i}\left[\Sigma^{-1}\left(\mu - \log P - r_f + \frac{1}{2}\operatorname{diag}(\Sigma)\right)\right]_i \tag{32}$$

This implies that when considering the relation between portfolio weights, $\omega_i = P_i D_i / W$, and log prices, the elasticity matrix is:

$$\mathcal{E} = \frac{\partial \omega}{\partial \log P} = -\frac{1}{\gamma}\Sigma^{-1}. \tag{33}$$

This is the same elasticity as the CARA case, albeit with different units, expenditure shares on log prices. Therefore, all our earlier discussion about the relation between the property of the covariance matrix and the identification assumptions apply to this case as well.

2.4.2 Logit

The logit model is commonly used in the industrial organization literature. There it is most often motivated by aggregation of a consumer discrete choice model, but can also apply to an individual choice of consumption basket.[9] While this model does not derive from standard optimal portfolio choice amont risky assets, its simplicity and tractability make it appealing for constructing empirical models (Koijen and Yogo, 2019).

[8] We log-linearize the return of portfolio ω, $r_p = \log R_p$ as:

$$r_p - r_f = \log\left(\omega' \exp\left(\mathbf{r} - r_f\right)\right) \simeq \omega'(\mathbf{r} - r_f) + \frac{1}{2}\omega' \operatorname{diag}(\Sigma) - \frac{1}{2}\omega' \Sigma \omega \tag{31}$$

[9] Anderson et al. (1988) derives the utility that leads to logit shares as optimal demand.

For an investor with initial wealth W, the expenditure shares or portfolio weights are:[10]

$$\omega_i = \frac{P_i D_i}{W} = \frac{\exp\left(-\alpha p_i + \theta' X_i + \epsilon_i\right)}{1 + \sum_l \exp\left(-\alpha p_l + \theta' X_l + \epsilon_l\right)}, \tag{35}$$

where p_i is the log of the price of asset i, X_i some observable characteristics, and ϵ_i unobservable characteristics.

When considering the relation between log portfolio weights and log prices, the elasticity is:

$$\mathcal{E} = \frac{\partial \log \omega}{\partial \log P} = -\alpha \left(\mathbf{I} - \mathbf{1}\omega'\right). \tag{36}$$

The coefficient α is the only demand parameter that determines the matrix of demand elasticity, as opposed to the whole covariance matrix in the CARA and CRRA cases. Further, this matrix always satisfies assumptions A1 ($\mathcal{E}_{ii} - \mathcal{E}_{ji} = \alpha$) and A2 ($\mathcal{E}_{jk} = \mathcal{E}_{ik} = \alpha\omega_k$), with α being the relative elasticity of demand.

3 Price Impact

We turn to a related empirical exercise, the estimation of price impact or multipliers. After setting up the corresponding regression framework, we show how our identification results apply to this situation. We then relate demand estimates and price impact estimates.

3.1 Price impact regression

Price impact measures how much prices change in response to an exogenous shift in demand. Of course, in equilibrium, realized aggregate demand cannot change because assets are in

[10]If there is not outside good, the model of expenditure shares becomes:

$$\omega_i = \frac{P_i D_i}{W} = \frac{\exp\left(-\alpha p_i + \theta' X_i + \epsilon_i\right)}{\sum_l \exp\left(-\alpha p_l + \theta' X_l + \epsilon_l\right)}. \tag{34}$$

fixed supply. However, it is possible for demand curves to shift. An idealized example would be an investor waking up in the morning and deciding to buy one share of Apple for no specific reason. In practice, empiricists have used shifts in asset purchases by central banks or rebalancing due to flows in and out mutual funds (Lou, 2012). Given such a shift in demand ΔD, causal inference estimates the multiplier $\widehat{\mathcal{M}}$ using the regression:

$$\Delta P_i = \widehat{\mathcal{M}} \Delta D_i + v_i. \tag{37}$$

There is no first-stage because the shift in demand curve is measured directly. However, there is still a stringent exclusion restriction: $\Delta D_i \perp v_i$. In words, the change in demand under consideration must be orthogonal to any other demand shift in the economy. For example, if a group of investors systematically mimicks the Fed's asset purchases, exogeneity is violated and the regression will be biased.

The same issue as for the estimation of demand elasticity arises: all prices are determined together in equilibrium. There is no such thing as the multiplier but instead a matrix $\mathcal{M}$ of own-demand and cross-demand multipliers:

$$\Delta P = \mathcal{M} \Delta D \tag{38}$$

It is immediate that all of the considerations discussed in Section 1 also apply to this relation. Specifically, the causal inference regression of equation (37) is well specified if the matrix $\mathcal{M}$ satisfies assumptions A1 and A2. In this case, the regression identifies the relative multiplier $\widehat{\mathcal{M}} = \mathcal{M}_{ii} - \mathcal{M}_{\text{cross}}(X_i, X_i)$. How much does the price of Ford change relative to the price of General Motors if demand for Ford changes relative to the demand for General Motors, where Ford and General Motors have the same observables.

3.2 Link with demand elasticity

Beyond the symmetry between the price impact regression and the demand elasticity regression, the two problems are intimately connected economically. Let us write the aggregate demand curve $D(P)$, the sum of the demand curves of all agents in the economy. The corresponding elasticity matrix is $\mathcal{E} = \partial D/\partial P$. In equilibrium, prices have to be such that aggregate demand equals the aggregate supply S, such that $D(P) = S$. If demand curves shift by an amount ΔD, the new equilibrium price $P + \Delta P$ satisfies $D(P + \Delta P) + \Delta D = S$. Using the implicit function theorem we immediately see that[11]

$$\mathcal{M} = -\left(\frac{\partial D}{\partial P}\right)^{-1} = -\mathcal{E}^{-1}. \tag{39}$$

The multiplier matrix is the inverse of the elasticity matrix. Of course inverting a matrix is different from inverting each of its elements. Thus, the own-price multiplier and the elasticity are not the inverse of each other ($\mathcal{M}_{ii} \neq -1/\mathcal{E}_{ii}$) and the same can be said of the cross-price multiplier and elasticity.

This raises the question whether the price impact regression reveals different information than the demand elasticity regression about the elasticity matrix $\mathcal{E}$. The next proposition states that this is not the case when the regressions are well specified.

Proposition 3 *If the elasticity matrix $\mathcal{E}$ satisfies assumptions A1 and A2, the multiplier matrix $\mathcal{M} = \mathcal{E}^{-1}$ satisfies them as well and the relative elasticity and multiplier are the inverse of each other:*

$$\widehat{\mathcal{M}} = -1/\widehat{\mathcal{E}} \tag{40}$$

The proposition has two parts, each shown in Appendix A.2. First, it states having a well-specified demand elasticity regression also implies that the price impact regression is

[11] We difference out the two equilibria (before and after the shift) to get $\frac{\partial D}{\partial P}\Delta P + \Delta D = 0$ which gives the multiplier as the effect of the change in demand on prices $\Delta P = -\left(\frac{\partial D}{\partial P}\right)^{-1}\Delta D$.

well-specified, and vice-versa. Second, under these conditions the relative multiplier coincides with the inverse of the relative elasticity despite neither of their individual component being stable by inversion: $\mathcal{M}_{ii} - \mathcal{M}_{\text{cross}}(X_i, X_i) = -\frac{1}{\mathcal{E}_{ii} - \mathcal{E}_{\text{cross}}(X_i, X_i)}$.

As we describe in Section 2.4, it is sometimes more suitable to estimate demand elasticities in different units (logarithms, portfolio shares instead of quantities, ...) The inversion result of Proposition 3 still applies to these different cases but with slightly adjusted formulas:

$$\mathcal{M}_{\{\log P, \log Q\}} = -\mathcal{E}^{-1}_{\{\log Q, \log P\}}, \tag{41}$$

$$\mathcal{M}_{\{\log P, \log Q\}} = -\left[\mathcal{E}_{\{\log \omega, \log P\}} - (\mathbf{I} - \mathbf{1}\omega')\right]^{-1}, \tag{42}$$

$$\mathcal{M}_{\{\log P, \log Q\}} = -\left[\text{diag}(\omega)^{-1}\mathcal{E}_{\{\omega, \log P\}} - (\mathbf{I} - \mathbf{1}\omega')\right]^{-1}. \tag{43}$$

For example in the case of logit where demand elasticity is measured by regressing the log portfolio share on log price, equation (42) gives us the multiplier in log units: by how many percents do prices move in response to a one percent change in aggregate demand. Similarly, equation (43) is useful for the case of CRRA.

3.3 Example: Relative multipliers in corporate bonds

We introduce an empirical example that we will follow throughout the rest of the paper. The example uses investment-grade corporate bonds between 2011Q2 to 2022Q3, which we aggregate to five maturity buckets: <3 years, 3–5 years, 5–7 years, 7–10 years, and >10 years, where cutoffs are chosen so that the number of observations per bucket is approximately equal. In total, we have $(N = 5) \times (T = 46) = 230$ observations. *Setup to be completed.*

Column (1) shows a significant relative multiplier when omitting any observables other than an overall intercept. Column (2) adds a date fixed effect. As discussed in section 2.3.1, under strong-form homogeneity in a comparable estimation sample the only observable for the cross-sectional regression is a cross-sectional constant, which absorbs substitution. Since this example is one of repeated cross-sections, each observable — which in specification (1)

Table 2: **Relative price multiplier $\widehat{\mathcal{M}}$ in corporate bonds**

	Price ΔP_{it}				
	(1)	(2)	(3)	(4)	(5)
Z_{it}	4.051*	0.995	**1.232**		
	(1.990)	(1.399)	(1.048)		
Z_{it}^{idio}				**1.232**	**1.232**
				(1.048)	(1.046)
Date Fixed Effects		Yes	Yes	Yes	Yes
X_i × Date Fixed Effects			Yes	Yes	
N	230	230	230	230	230
R^2	0.064	0.724	0.981	0.981	0.724

Table 2 reports the results of price-multiplier regressions for investment-grade corporate bonds across maturity buckets. Bold estimates correspond to our preferred specifications accounting for duration-based substitution. The sample period is 2011Q2 to 2022Q3. Standard errors are clustered by date.

is again just an intercept — should be interacted with a date fixed effect. Accounting for such homogeneous substitution already matters; it reduces the relative multiplier from about 4 to 1. Column (3) is our preferred specification. It adds duration-based substitution by interacting the normalized duration characteristic, X_i, with date fixed effects. Accounting for duration-based substitution slightly increases the relative multiplier to $\widehat{\mathcal{M}} = 1.2$. Overall these results are consistent with estimates of $\widehat{\mathcal{M}}$ from the literature, specifically Chaudhary et al. (2022). Columns (4) and (5) provide alternative specifications that identify the same relative multiplier as column (3). Specifically, the orthogonalized instrument Z_{it}^{idio} is by construction orthogonal to the observable, so that as an application of the Frisch-Waugh-Lovell theorem $X_i \times$ Date Fixed Effects may be omitted and still the same relative multiplier is identified from the regression.

4 Estimating Substitution Patterns and The Aggregate Elasticity

We now ask how to estimate the remainder of the elasticity matrix. We first review a few interesting counterfactual questions about financial markets, and show which ones can be answered using the cross-sectional causal estimates alone. Then we show which empirical strategies are necessary to complement these estimates in order to tackle the remaining questions.

4.1 Interesting Counterfactuals and the Limits of the Cross-Section

The matrix of demand elasticity $\mathcal{E}$ and its counterpart the multiplier $\mathcal{M}$ are key to conduct counterfactual experiments. What are the overall effects of a shift in investor demand? Once we move from questions about the relative price adjustment of different stocks that are affected by the demand, to questions about the average level of prices, relying solely on our relative estimates ($\hat{\mathcal{E}}$ or $\hat{\mathcal{M}}$) is no longer sufficient.

We have already broached questions of relevance to policy makers such as what happens when a specific financial institution is in trouble and is being forced to liquidate its portfolio. Two significant examples highlight the relevance of identifying if not the whole matrix of multiplier $\mathcal{M}$ at least more than the relative multiplier (or elasticity). First consider the impact of broad shifts in the demand for ESG affect equilibrium prices, the reclassification of specific stocks from green to brown change the price of these stocks. Second, what are the effects of a central bank changing the composition of its bond portfolio (operation twist)? What are the effects of a central bank' decision to purchase a large quantity of bonds with a specific duration?

We present decomposition results where $\hat{\mathcal{M}}$ represents only a small part of the response — the part related to relative price movements. We characterize the components of the multiplier matrix $\mathcal{M}$ which are relevant to answer these more aggregated questions accurately.

First, we describe the limits of the cross-section to answer these questions in the context of a multiplier that is fully symmetric across assets. This simple case highlight the necessity for another source of variation in demand to identify the answer to these more aggregated questions.

Specifically, we start with a multiplier matrix that is such that each asset share the same response to its own demand and to other's demand as follows:

$$\mathcal{M} = \mathcal{M}_{\text{own}}\mathbf{I} + (\mathcal{M}_{\text{cross}} - \mathcal{M}_{\text{own}})\,\mathbf{1}\mathbf{1}' = \begin{bmatrix} \mathcal{M}_{\text{own}} & \mathcal{M}_{\text{cross}} & \cdots & \mathcal{M}_{\text{cross}} \\ \mathcal{M}_{\text{cross}} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathcal{M}_{\text{cross}} \\ \mathcal{M}_{\text{cross}} & \cdots & \mathcal{M}_{\text{cross}} & \mathcal{M}_{\text{own}} \end{bmatrix}$$

Looking at the change in the demand of asset 1, while asset 2 is unchanged the cross-sectional difference will only inform the researcher about $\widehat{\mathcal{M}}$:

$$\varDelta P_1 - \varDelta P_2 = \mathcal{M}_{\text{own}}\varDelta D_1 - \mathcal{M}_{\text{cross}}\varDelta D_1 = \widehat{\mathcal{M}}\varDelta D_1$$

If the researcher is interested in the estimates of the effect of an hypothetical uniform shift in demand, such that for all assets i $\varDelta D_i = \varDelta D$, the effect will give find the following relation

$$\varDelta P_1 = (\mathcal{M}_{\text{own}} + (N-1)\mathcal{M}_{\text{cross}})\,\varDelta D = \left(\widehat{\mathcal{M}} + N\mathcal{M}_{\text{cross}}\right)\varDelta D$$

Thus, the sole part of $\mathcal{M}$ that the researcher can estimate from the relative differences in Section 2 are not enough to evaluate the effect of $\varDelta D$ on the change in the price $\varDelta P_1$ which depend on both $\mathcal{M}_{\text{own}}$ and $\mathcal{M}_{\text{cross}}$ separately. Making matters more challenging, in presence of excluded assets or with grouping of assets, the cross-elasticities from these outside assets become relevant but in the cross-sectional regression they are absorbed in the group fixed effects.

This problem is the classic problem of aggregation from micro-level estimates to aggregate counterfactual, sometimes known as the missing-intercept problem. To make further progress on this issue, we follow two different but non-exclusive avenues. One path is to make stronger structural assumptions, i.e. restrictions on the multiplier matrix, where the local multiplier is informative for the more aggregate multiplier; this is akin to using portable moments in macroeconomics to discipline structural model (Nakamura and Steinsson, 2018). Alternatively, one can find new and specific sources of variation in the data to inform us on the relevant characteristics of the multiplier matrix $\mathcal{M}$. While we mostly focus on the latter, we show how some of the structural assumptions from empirical researchers complement the search for new instruments.

4.2 From micro to macro: a decomposition in a simple case

Our first aggregation result presents the case of a single homogeneous group of assets such that our assumption A1 is satisfied within the group. In this case the elastic matrix does not have to present any symmetry, yet we present a similar decomposition of the multiplier between an aggregate multiplier and a relative multiplier. The aggregate multiplier can only be estimated using variation from demand in the time series.

Specifically, we start from a multiplier matrix that satisfies A1.[12] Under these conditions, there is a specific combination of changes in all of the demand which we call aggregate demand shock $\varDelta D_{\text{agg}} = \theta' \varDelta D$, with $\theta' \mathbf{1} = 1$, that leads to an uniform response in prices $\varDelta P_{\text{agg}} = \theta' \varDelta P$. The overall response of the price to changes in demand can be decomposed into this aggregate response and a change in the relative price. We state the result formally:

Proposition 4 (Multiplier Decomposition for a Homogeneous Group) *Take a matrix $\mathcal{M}$ satisfying assumption A1 and vector of diagonal elements $\mathcal{M}_{own}$ as given. Then, there exists a set of weights θ, with $\theta' \mathbf{1} = 1$, such that if we define the aggregate change in*

[12] In this Section we use as an example for our result, the multiplier regression which tries to estimate $\mathcal{M}$. All of the results are also valid for the demand elasticity regression and $\mathcal{E}$.

price $\Delta P_{agg} = \theta' \Delta P$ and in demand $\Delta D_{agg} = \theta' \Delta D$, we have:

$$\Delta P = \left(\bar{\mathcal{M}} \Delta D_{agg}\right) \cdot \mathbf{1} + \widehat{\mathcal{M}}(\Delta D - \Delta D_{agg} \mathbf{1}) \tag{44}$$

$$\Delta P_{agg} = \bar{\mathcal{M}} \Delta D_{agg}. \tag{45}$$

Moreover the aggregation weights are $\theta = (\mathcal{M}_{own} - \widehat{\mathcal{M}}\mathbf{1})/[(\mathcal{M}_{own} - \widehat{\mathcal{M}}\mathbf{1})'\mathbf{1}]$ and the aggregate uniform multiplier is $\bar{\mathcal{M}} = \mathcal{M}_{own}'\mathbf{1} - (N-1)\widehat{\mathcal{M}}$.

The proof is in Appendix A.4. Note that the weights θ that define the aggregate demand shock have a natural interpretation for the cases developed above with a few structural assumptions. First if $\mathcal{M}$ is symmetric (or equivalently $\mathcal{M}_{own}$ is constant), then aggregation corresponds to a simple average $\theta = N^{-1}\mathbf{1}$. Furthermore $\bar{\mathcal{M}} = \hat{\mathcal{M}} + N\mathcal{M}_{\text{cross}}$, as we recover the case from Section 4.1. In the case of logit, the aggregation weights assets are proportional to their portfolio share $\theta = \omega/(\omega'\mathbf{1})$. In this case $\bar{\mathcal{M}} = -\alpha w_0$, where w_0 is the portfolio weight on the outside asset. Intuitively, the aggregate shock is driven by the stocks that figure more prominently in terms of cross-substitution, i.e. that has more weights on the effect of all the other shocks: in the case of logit this maps exactly to the portfolio shares.

Finally, note that to estimate the aggregate effect it is not necessary to estimate the whole matrix of multiplier $\mathcal{M}$. It is enough to find variation in ΔD_{agg} to estimate the aggregate multiplier $\bar{\mathcal{M}}$. This simple case only requires one source of variation in the time series to fully decompose micro and macro-multipliers. We now turn to two cases where the set of asset is no longer homogeneous.

4.3 Estimating observable-based multipliers

Outside the simple case where substitution is unconditionally homogeneous and where there is only one notion of aggregate price multiplier — the macro multiplier — there are potentially other, observable-based aggregators driving substitution and price impact. Specifically, in the context of Proposition 1, the observables that drive substitution, and need to be conditioned on to identify $\widehat{\mathcal{E}}$ from the cross-section, are the same observables that potentially matter for aggregate questions.

For exposition, consider a combination of aggregate substitution as with unconditional homogeneity with a single observable X, that varies in the cross-section, and that drives heterogeneous substitution patters. Proposition 5 shows how to estimate the $\mathcal{M}_S$, the matrix of aggregate multipliers along the observable X and the aggregate. For the proof, see Appendix A.3.

Proposition 5 (Observable-based multiplier regression) *Take a matrix $\mathcal{M}$ satisfying assumptions A1 and A2 and that can be written as*

$$\mathcal{M} = \widehat{\mathcal{M}}\mathbf{I} + \left[\frac{\mathbf{1}}{\mathbf{1}'\mathbf{1}}, X\right] \mathcal{M}_S \left[\frac{\mathbf{1}}{\mathbf{1}'\mathbf{1}}, X\right]', \tag{46}$$

where X is such that $1'X = 0$ and $X'X = 1$. Further, assume that the demand shifter $Z \perp \{u|X\mathbf{1}_t, \mathbf{1}_t\}$; within each period variation in Z that is orthogonal to X is orthogonal to variation in u. Then, the matrix $\mathcal{M}_S$ can be estimated from aggregated time-series regressions:

$$\mathbb{E}\left[\Delta\tilde{P}_{agg}|Z_{agg}, Z_X\right] = \mathcal{M}_{S[1,1]}Z_{agg} + \mathcal{M}_{S,[1,2]}Z_X \tag{47}$$

$$\mathbb{E}\left[\Delta\tilde{P}_X|Z_{agg}, Z_X\right] = \mathcal{M}_{S,[2,1]}Z_{agg} + \mathcal{M}_{S,[2,2]}Z_X \tag{48}$$

where $\Delta\tilde{P}_{agg} = \mathbf{1}'\left(\Delta P - \widehat{M}Z\right)$, $\Delta\tilde{P}_X = X'\left(\Delta P - \widehat{M}Z\right)$, $Z_{agg} = \frac{\mathbf{1}'Z}{\mathbf{1}'\mathbf{1}}$, and $Z_X = X'Z$.

Proposition 5 shows how to estimate the substitution multiplier matrix $\mathcal{M}_S$ from aggregate time-series regressions of aggregate price changes and observable-weighted aggregate price changes on an aggregate time-series instrument and an aggregate observable-tilted time-series instrument. Like in the case of just estimating the macro multiplier in section 4.2, the cross-section does not provide a way of estimating aggregate multipliers, only the time-series does. This is a version of the missing intercept problem for cross-sections, only that here there are multiple missing intercepts.

An implication of Proposition 5 is that if one wants to estimate the overall macro-multiplier, corresponding to $\mathcal{M}_{S[1,1]}$ in equation (47), doing so in isolation from a univariate regression of $\Delta\tilde{P}_{agg}$ on Z_{agg} will, in general, not correctly identify $\mathcal{M}_{S[1,1]}$. This is because Z_{agg} and Z_X are generally not orthogonal, and hence the univariate regression will suffer from omitted variable bias. More generally, with more than one observable driving aggregate substitution, all dimensions of aggregate substitution need to be included in the regression. Huber (2023) highlights a similar problem in the context of general equilibrium spillovers. We direct a further discussion of the implications of this result for doing empiricial work to the next subsection.

Beyond the macro multiplier, Proposition 5 shows how to estimate the other elements of $\mathcal{M}_S$, which enable answering additional questions. Equation (48) estimates how price multipliers differ along the characteristic X; $\mathcal{M}_{S,[2,2]}$ tells us how much more the price of the group of assets with high X moves relative to that of the group of assets with low X in response to an aggregate shock ($\mathcal{M}_{S,[2,1]}$) or an observable-tilted shock ($\mathcal{M}_{S,[2,2]}$). In particular, $\mathcal{M}_{S,[2,1]}$ tells us how the macro multiplier varies in the cross-section of observables. For example, in the context of treasury bonds, duration is a key observable that may affect the magnitude of the macro multiplier across maturities. In particular, it is natural to think that an aggregate shock, such as the Federal reserve stepping in and equally purchasing bonds of all maturities, will likely move the price of long-term bonds more so than that of short-term bonds.

$\mathcal{M}_{S,[1,2]}$ in (47) measures how shocks tilted along a certain observable may affect aggregate prices. Consider a variation on the previous example where the Federal Reserve purchases long-term bonds but sells short-term bonds: "operation twist". Even though such an operation would not constitute any net buying of treasuries, it may still affect overall bond prices because it does correspond to a large duration-weighted shock. Note how this is different from the Federal Reserve deciding to randomly purchase one bond over another (e.g., Selgrad, 2023). In that case, the shock does not systematically line up with any dimension of aggregate substitution, and hence the relative multiplier $\widehat{\mathcal{M}}$ will provide the answer to how much more the treated bond's price moves relative to that of the control's.

4.3.1 Practical advice

The result from Proposition 5 may appear depressing: to estimate, for example, the macro multiplier, from a regression of aggregate price changes on an aggregate shock, the econometrician needs to control for all other aggregated shocks along any observable that drives aggregate substitution patterns. Otherwise, the econometrician will risk introducing omitted variable bias, unless the aggregate shock Z_{agg} used to estimate the macro multiplier is orthogonal to all other, observable-based aggregate shocks Z_X.

In practice, it is unlikely that the econometrician will be able to incorporate, or even know, every possible dimension of aggregate substitution. And since for aggregate multipliers, the econometrician can only lean into variation from the time series, they realistically will only have a sufficiently long sample to incorporate that many observables while still maintaining sufficient statistical power. As a result, the econometrician will need to take a stand on which dimensions of substitution to include in their empirical design.

We offer some guidance: First, the econometrician should naturally include dimensions of substitution that are likely to be important for the research question at hand; in particular, if it is variation along these dimensions that they seek to explain. For example, a researcher interested in studying how high-yield versus investment-grade corporate bond prices move differentially relative to an aggregate shock will want to make sure that their aggregate shock is either orthogonal to shocks disproportionally affecting high-yield verus investment-grade bonds, or explicitly control for such credit shocks.

Second, the econometrician should include substitution along dimensions that they believe, as perhaps grounded in theory, are associated with substantial risk premia. For example, in the study of treasuries, many theories predict that bond duration is a key driver of risk manifesting itself in term premia. So naturally, irrespective of the precise research question, a researcher studying treasury prices or yield curves should consider including duration-based substitution because it is a ubiquitous and theoretically well-grounded dimension of substitution in the treasury market.

And third, in the study of macro multipliers, we encourage researchers to open up the black box of aggregate demand shocks to look at whether the aggregate shock is tilted along any dimension. Even though the cross-section does not identify anything about aggregate elasticities — the missing intercept problem — it can tell us something about whether an aggregate shock is really a parallel shift for the entire cross-section or varies systematically along some dimension. Because if it does vary systematically in the cross-section, the regression may not identify the macro multiplier, but a quantity also related to observable-based substitution.

4.3.2 Example: Duration-based multipliers in corporate bonds

Table 3 continues the example of price multiplier regressions from section 3.3 for 5 maturity buckets in investment-grade corporate bonds, and reports the entire estimated price multiplier matrix $\mathcal{M}$.

Specifically, specifications (3) and (4) correspond to equations (47) and (48) in Proposition 5, respectively. Put together they characterize the observable-based multiplier matrix $\mathcal{M}_S$ in its entirety. The overall macro multiplier $\mathcal{M}_{S[1,1]}$ is estimated to be around 5.64. Notably, the estimated macro multiplier varies with the cross-section of duration X, as can also be seen non-parametrically from the left panel of Figure 2. There may also be a role for "operation twist"-like shocks to high-low duration bonds in affecting aggregate bond prices, even though the pattern is not entirely monotonic when estimated non-parametrically in the right panel of Figure 2.

Specifications (1) and (2) show how to estimate the observable-based multiplier matrix $\mathcal{M}_S$ from a panel. Importantly, the additional cross-sectional variation does not help with

Table 3: **Estimating the entire multiplier matrix $\mathcal{M}$ in corporate bonds**

	Price ΔP_{it}		Aggregate Price ΔP_t^{agg}	Factor Price ΔP_t^X
	(1)	(2)	(3)	(4)
Z_{it}^{idio}	1.232	1.232		
	(1.053)	(1.058)		
X_i	0.005	0.007		
	(0.005)	(0.005)		
Z_t^{agg}	5.640*	5.640*	5.640	8.154*
	(2.768)	(2.780)	(2.807)	(3.135)
Z_t^X	0.886	0.886	0.886	4.690
	(3.666)	(3.682)	(3.717)	(3.751)
$Z_t^{agg} \times X_i$		8.154*		
		(3.105)		
$Z_t^X \times X_i$		4.690		
		(3.716)		
N	230	230	46	46
R^2	0.085	0.116	0.109	0.120

Table 3 reports the results of estimating the entire multiplier matrix $\mathcal{M}$ in investment-grade corporate bonds with five maturity buckets. Specfciations (3) and (4) correspond to equations (47) and (48) from Proposition 5 and estimate the four elements of the observable-based $\mathcal{M}_S$ from aggregated time-series regressions. Specifications (1) and (2) accomplish the same based on panel regressions, and additionally identify the relative multiplier $\widehat{\mathcal{M}}$. The sample period is 2011Q2 to 2022Q3. Standard errors are clustered by date.

identifying aggregate multipliers — the missing intercepts problem — but identify the exact same coefficients as from the aggregated time-series regressions. This highlights again that the variation in the panel that identifies aggregate multipliers comes entirely from the time-series. The cross-section can only recover the relative multiplier $\widehat{\mathcal{M}}$, which matches the estimate of the relative multiplier from section 3.3.

4.4 A general decomposition with multiple groups

In this section, we move away from patterns of substitutions that are driven by factors which slice across groups of assets. We consider the case of multiple groups of homogeneous assets.

There are G groups of assets, and each group g include a number N_g of assets such the

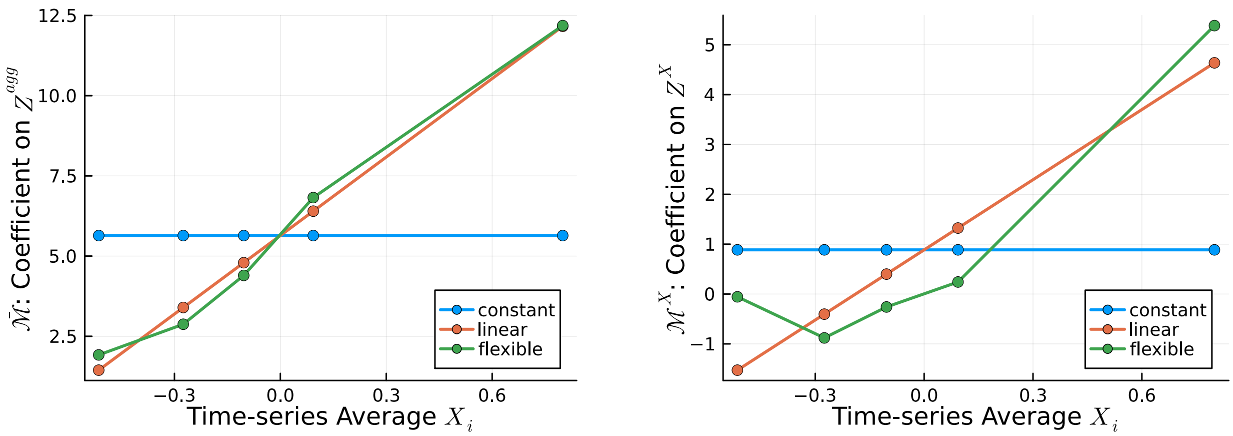

Figure 2: **Duration-based price multipliers in corporate bonds.** Figure 2 puts together the results from Table 3 to show how the macro-multiplier varies in the cross-section of corporate bond duration (left panel), and how asymmetric shocks to high-low duration bonds differentially affect bond prices of high- versus low duration bonds (right panel). The blue line is a constant, corresponding to the unconditional estimated macro multiplier $\mathcal{M}_{S[1,1]}$ from specification (1) in Table 3 on the left, and the unconditional multiplier from an "operation twist"-type shock on the right. The orange line shows how these multipliers vary linearly with duration, and the green line shows these multipliers for every single maturity bucket. The sample period is 2011Q2 to 2022Q3.

multiplier matrix $\mathcal{M}$ can be define by its group-blocks as:

$$\mathcal{M}_{gh}^{\text{group}} = [\mathcal{M}_{ij}]_{\{i\in g, j\in h\}} \tag{49}$$

From Section 4.2, we already have a meso-decomposition of the diagonal blocks $\mathcal{M}_{gg}^{\text{group}}$ between the group-level response and the relative response within the group.

Proposition 6 (Multiplier Decomposition for Multiple Groups) *Take a matrix $\mathcal{M}$ composed of G groups of assets such that within a group assumption A1 and A2 are satisfied. Then, from the single group decomposition of proposition 4, we obtain the group aggregates for each group g, $\Delta D_{group,g}$ and $\Delta P_{group,g}$. There exists a set of group-level weights $\boldsymbol{\vartheta} = \{\vartheta_{gh}\}_{g\in G, h\in G}$, such that if we define the aggregate change in price $\Delta P_{agg} = \vartheta' \Delta P_{group}$ and in demand $\Delta D_{agg} = \vartheta' \Delta D_{group}$, we have*

$$\Delta P_g = \widehat{\mathcal{M}} \Delta D_{idio,g} + \bar{\mathcal{M}}_{gg} \Delta D_{agg,g} + \sum_{h\neq g} \bar{\mathcal{M}}_{gh} \vartheta'_{hh} \Delta D_h \mathbf{1} + \sum_{h\neq g} \mathbf{1} \xi'_{gh} \Delta D_h, \tag{50}$$

where the aggregation weights are defined from $\mathcal{M}_{gh} = \mathbf{1}\vartheta_{gh}$, the aggregated group-level effects are defined as $\bar{\mathcal{M}}_{gh} = (\vartheta'_{hh}\vartheta_{hh})^{-1}\vartheta_{hh}\vartheta_{gh}$, and the residual cross-substitution is $\xi_{gh} = \vartheta_{gh} - \bar{\mathcal{M}}_{gh}\vartheta_{hh}$.

Moreover if we assume that $\xi_{gh} = 0$, we have a well-defined aggregated system:

$$\Delta P_{agg} = \bar{\mathcal{M}}\Delta D_{agg}. \tag{51}$$

More generally we write the decomposition:

$$\mathcal{M}_{gh} = \bar{\mathcal{M}}_{gh}N_h^{-1}\mathbf{1}\mathbf{1}' + \mathcal{M}_{within\ group\ spillover} \tag{52}$$

$$\bar{\mathcal{M}}_{gh} = [\mathcal{M}_{gh}]_{1\bullet}\mathbf{1}, \qquad \textit{sum of the rows for } \mathcal{M}_{gh} \tag{53}$$

$$\mathcal{M}_{within\ group\ spillover} = \mathbf{1}\vartheta'_{gh}, \qquad \textit{with } \vartheta'_{gh}\mathbf{1} = 0. \tag{54}$$

This general case covers cases that seek to relax the homogeneity assumption. Of course the symmetric case still features equal weights of the aggregator such that we have $\vartheta_{hh} = N_h^{-1}\mathbf{1}$, and the off-diagonal aggregated elasticity terms are $\bar{\mathcal{M}}_{gh} = N_h \left[\mathcal{M}_{gh}\right]_{i\in G, j\in H}$.

The nested logit structure relaxes the IIA assumption inherent to logit and allows for more flexibility in the substitution across different types of assets and fits into the group decomposition.

5 Conclusion

This paper provides a framework fo using causal inference in the asset pricing context. Specifically, we provide conditions for valid estimation in presence of the natural spillovers that exist between assets when making portfolio choices. The two conditions are constant relative elasticity and homogenous substitution conditional of observable. The latter implies that two assets are comparable if the demand for them responds in the same way to the price of every other assets. We provide guidelines to design experiment satisfying these conditions and assess their plausibility in the data. We also show that they map to restrictions often imposed in standard asset pricing models. When these conditions hold, the standard cross-sectional difference-in-difference or instrumental variable approach identifies the relative elasticity between comparable assets, the difference between their own-price and cross-price elasticity. Other dimensions of substitutions such as separating own-price and cross-price elasticity, the macro elasticity, or responses to shocks across broad categories of assets, are jointly estimated by a set of time series regressions. These simple tools and principles offer a straightforward package for researchers wanting to use natural experiments to better understand investment decisions and their equilibrium impact.

References

Anderson, Simon P, André De Palma, and J-F Thisse, "A representative consumer theory of the logit model," *International Economic Review*, 1988, pp. 461–466.

Angrist, Joshua D. and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2009.

Campbell, John Y., *Financial Decisions and Markets: A Course in Asset Pricing*, Princeton University Press, 2017.

_ **and Luis Viceira**, *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors*, first ed., Oxford, UK: Oxford University Press, 2002.

Chang, Yen-Cheng, Harrison Hong, and Inessa Liskovich, "Regression Discontinuity and the Price Effects of Stock Market Indexing," *The Review of Financial Studies*, 07 2014, *28* (1), 212–246.

Chaudhary, Manav, Zhiyu Fu, and Jian Li, "Corporate bond multipliers: Substitutes matter," Technical Report 2022.

Cochrane, John H, *Asset pricing*, Princeton University Press, 2005.

Coppola, Antonio, "In safe hands: The financial and real impact of investor composition over the credit cycle," Technical Report, Working paper 2021.

Deaton, Angus and John Muellbauer, "An almost ideal demand system," *The American economic review*, 1980, *70* (3), 312–326.

Fuchs, William, Satoshi Fukuda, and Daniel Neuhann, "Demand-System Asset Pricing: Theoretical Foundations," Technical Report 2024.

Gabaix, Xavier and Ralph SJ Koijen, "In search of the origins of financial fluctuations: The inelastic markets hypothesis," Technical Report, National Bureau of Economic Research 2021.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár, "Contamination bias in linear regressions," Technical Report, National Bureau of Economic Research 2022.

Haddad, Valentin, Paul Huebner, and Erik Loualiche, "How competitive is the stock market? theory, evidence from portfolios, and implications for the rise of passive investing," *Working paper*, 2024.

Huber, Kilian, "Estimating general equilibrium spillovers of large-scale shocks," *The Review of Financial Studies*, 2023, *36* (4), 1548–1584.

Koijen, Ralph S. J. and Motohiro Yogo, "A Demand System Approach to Asset Pricing," *Journal of Political Economy*, 2019, *127* (4), 1475–1515.

Lou, Dong, "A flow-based explanation for return predictability," *The Review of Financial Studies*, 2012, *25* (12), 3457–3489.

Markowitz, Harry M, "Portfolio selection," *The journal of finance*, 1952, *7*, 77–91.

Nakamura, Emi and Jón Steinsson, "Identification in Macroeconomics," *Journal of Economic Perspectives*, August 2018, *32* (3), 59–86.

Selgrad, Julia, "Testing the Portfolio Rebalancing Channel of Quantitative Easing," Technical Report 2023.

Shleifer, Andrei, "Do Demand Curves for Stocks Slope Down?," *The Journal of Finance*, 1986, *41* (3), 579–590.

Appendix

A Proofs

A.1 Proof of Proposition 2

The data-generating process under heterogenous treatment effects is:

$$\mathit{\Delta} D_i = \mathcal{E}_{ii} \mathit{\Delta} P_i + \sum_{j \neq i} \mathcal{E}_{ij} \mathit{\Delta} P_j + \epsilon_i \tag{55}$$

$$\mathit{\Delta} P_i = \lambda_i Z_i + u_i \tag{56}$$

The instrument Z_i, with constant variance $var(Z_i) = var(Z), \forall i$, is randomly assigned and independent of everything else:

$$Z_i \perp\!\!\!\perp Z_j \quad \forall i \neq j \tag{57}$$

$$Z_i \perp\!\!\!\perp \mathcal{E}_{kl} \quad \forall i, k, l \tag{58}$$

$$Z_i \perp\!\!\!\perp \lambda_j \quad \forall i, j \tag{59}$$

$$Z_i \perp\!\!\!\perp u_j \quad \forall i, j \tag{60}$$

$$Z_i \perp\!\!\!\perp \epsilon_j \quad \forall i, j \tag{61}$$

After substituting (56) into (55), we derive the estimate from the demand equation

$$\mathit{\Delta} D_i = \mathcal{E}_{ii} \lambda_i Z_i + \sum_{j \neq i} \mathcal{E}_{ij} \lambda_j Z_j + \mathcal{E}_{ii} u_i + \sum_{j \neq i} \mathcal{E}_{ij} u_j + \epsilon_i \tag{62}$$

Definitions and preliminiaries. Without loss of generality, define a centered instrument $\tilde{Z}_i$ as

$$\tilde{Z}_i \equiv Z_i - \frac{1}{N} \sum_j Z_j, \tag{63}$$

such that we have the following properties:

$$\sum_{j \neq i} \tilde{Z}_j = -\tilde{Z}_i \tag{64}$$

$$cov(\tilde{Z}_i, \tilde{Z}_j) = \underbrace{cov(Z_i, Z_j)}_{=0} - \frac{1}{N} \underbrace{\sum_k cov(Z_k, Z_j)}_{=var(Z)} - \frac{1}{N} \underbrace{\sum_l cov(Z_i, Z_l)}_{=var(Z)} + \frac{1}{N^2} \underbrace{cov(\sum_k Z_k, \sum_l Z_l)}_{=Nvar(Z)} \tag{65}$$

$$= -\frac{1}{N} var(Z) \tag{66}$$

Next, define $\bar{\lambda}$ and $\tilde{\lambda}_i$ such that:

$$\lambda_i = \bar{\lambda} + \tilde{\lambda}_i \tag{67}$$

$$\sum_j \lambda_j = N\bar{\lambda} \tag{68}$$

$$\sum_{j \neq i} \tilde{\lambda}_j = -\tilde{\lambda}_i \tag{69}$$

Finally, define $\mathcal{E}_{i,cross}$ as the λ_j weighted average of $\mathcal{E}_{ij}$:

$$\mathcal{E}_{i,cross} = \frac{\sum_{j \neq i} \lambda_j \mathcal{E}_{ij}}{\sum_{j \neq i} \lambda_j} \tag{70}$$

Proof. Based on the definitions above, rewrite $\sum_{j \neq i} \mathcal{E}_{ij} \lambda_j \tilde{Z}_j$ from equation (62) as:

$$\sum_{j \neq i} \mathcal{E}_{ij} \lambda_j \tilde{Z}_j = \mathcal{E}_{i,cross} \sum_{j \neq i} \lambda_j \tilde{Z}_j + \sum_{j \neq i} (\mathcal{E}_{ij} - \mathcal{E}_{i,cross}) \lambda_j \tilde{Z}_j \tag{71}$$

$$= \mathcal{E}_{i,cross} \bar{\lambda} \underbrace{\sum_{j \neq i} \tilde{Z}_j}_{=-\tilde{Z}_i} + \mathcal{E}_{i,cross} \sum_{j \neq i} \tilde{\lambda}_j \tilde{Z}_j + \sum_{j \neq i} (\mathcal{E}_{ij} - \mathcal{E}_{i,cross}) \lambda_j \tilde{Z}_j \tag{72}$$

$$= -\mathcal{E}_{i,cross} \bar{\lambda} \tilde{Z}_i + \mathcal{E}_{i,cross} \sum_{j \neq i} \tilde{\lambda}_j \tilde{Z}_j + \sum_{j \neq i} (\mathcal{E}_{ij} - \mathcal{E}_{i,cross}) \lambda_j \tilde{Z}_j \tag{73}$$

Plugging into equation (62):

$$\Delta D_i = \left(\mathcal{E}_{ii} \lambda_i - \mathcal{E}_{i,cross} \bar{\lambda} \right) \tilde{Z}_i + \mathcal{E}_{i,cross} \sum_{j \neq i} \tilde{\lambda}_j \tilde{Z}_j + \sum_{j \neq i} (\mathcal{E}_{ij} - \mathcal{E}_{i,cross}) \lambda_j \tilde{Z}_j + \mathcal{E}_{ii} u_i + \sum_{j \neq i} \mathcal{E}_{ij} u_j + \epsilon_i \tag{74}$$

We are interested in $cov(\Delta D_i, \tilde{Z}_i)$ and $cov(\Delta P_i, \tilde{Z}_i)$. Since $\tilde{Z}_i$ is mean-zero, by the law of iterated expectations we have:

$$cov(\Delta D_i, \tilde{Z}_i) = \mathbb{E}\left[\Delta D_i \tilde{Z}_i\right] = \mathbb{E}\left[\mathbb{E}\left[\Delta D_i \tilde{Z}_i | \Theta\right]\right], \tag{75}$$

where Θ is a set that contains all $\mathcal{E}_{ij}$ and λ_i.

We have:

$$\begin{aligned}
\mathbb{E}\left[\left(\mathcal{E}_{ii}\lambda_i - \mathcal{E}_{i,cross}\bar{\lambda}\right)\tilde{Z}_i^2 | \Theta\right] &= \left(\mathcal{E}_{ii}\lambda_i - \mathcal{E}_{i,cross}\bar{\lambda}\right) var(\tilde{Z}) && (76)\\
\mathbb{E}\left[\mathcal{E}_{i,cross}\sum_{j\neq i}\tilde{\lambda}_j \tilde{Z}_i \tilde{Z}_j | \Theta\right] &= \mathcal{E}_{i,cross}\sum_{j\neq i}\tilde{\lambda}_j \mathbb{E}\left[\tilde{Z}_i, \tilde{Z}_j\right] && (77)\\
&= -\frac{var(Z)}{N}\mathcal{E}_{i,cross}\sum_{j\neq i}\tilde{\lambda}_j && (78)\\
&= \frac{var(Z)}{N}\mathcal{E}_{i,cross}\tilde{\lambda}_i && (79)\\
&= \frac{N var(\tilde{Z})}{(N-1)^2}\mathcal{E}_{i,cross}\tilde{\lambda}_i && (80)\\
\mathbb{E}\left[\sum_{j\neq i}(\mathcal{E}_{ij} - \mathcal{E}_{i,cross})\lambda_j \tilde{Z}_i \tilde{Z}_j | \Theta\right] &= \sum_{j\neq i}(\mathcal{E}_{ij} - \mathcal{E}_{i,cross})\lambda_j \mathbb{E}\left[\tilde{Z}_i, \tilde{Z}_j\right] && (81)\\
&= -\frac{var(Z)}{N}\sum_{j\neq i}(\mathcal{E}_{ij} - \mathcal{E}_{i,cross})\lambda_j && (82)\\
&= -\frac{var(Z)}{N}\left(\frac{\sum_{j\neq i}\lambda_j \mathcal{E}_{ij}}{\sum_{j\neq i}\lambda_j}\sum_{j\neq i}\lambda_j - \mathcal{E}_{i,cross}\sum_{j\neq i}\lambda_j\right) && (83)\\
&= -\frac{var(Z)}{N}\left(\mathcal{E}_{i,cross}\sum_{j\neq i}\lambda_j - \mathcal{E}_{i,cross}\sum_{j\neq i}\lambda_j\right) && (84)\\
&= 0 && (85)\\
\mathbb{E}\left[\mathcal{E}_{ii}\tilde{Z}_i u_i + \sum_{j\neq i}\mathcal{E}_{ij}\tilde{Z}_i u_j + \tilde{Z}_i \epsilon_i | \Theta\right] &= 0 && (86)\\
\mathbb{E}\left[\Delta P_i \tilde{Z}_i | \Theta\right] &= \lambda_i var(\tilde{Z}) && (87)
\end{aligned}$$

Then:

$$cov\left(\mathit{\Delta} P_i, \tilde{Z}_i\right) = \mathbb{E}\left[\mathbb{E}\left[\mathit{\Delta} P_i \tilde{Z}_i | \mathit{\Theta}\right]\right] = \mathbb{E}\left[\lambda_i var(\tilde{Z})\right] = var(\tilde{Z})\mathbb{E}\left[\lambda_i\right] \tag{88}$$

$$cov\left(\mathit{\Delta} D_i, \tilde{Z}_i\right) = \mathbb{E}\left[\mathbb{E}\left[\mathit{\Delta} D_i \tilde{Z}_i | \mathit{\Theta}\right]\right] \tag{89}$$

$$= var(\tilde{Z})\left(\mathbb{E}\left[\mathcal{E}_{ii}\lambda_i\right] - \mathbb{E}\left[\lambda_i\right]\mathbb{E}\left[\mathcal{E}_{i,cross}\right] + \frac{1}{(N-1)^2}\mathbb{E}\left[\mathcal{E}_{i,cross}\tilde{\lambda}_i\right]\right) \tag{90}$$

The instrumental variable regression with heterogenous treatment effects identifies:

$$\widehat{\mathcal{E}} = \frac{cov\left(\mathit{\Delta} D_i, \tilde{Z}_i\right)}{cov\left(\mathit{\Delta} P_i, \tilde{Z}_i\right)} \tag{91}$$

$$= \frac{\mathbb{E}\left[\lambda_i \mathcal{E}_{ii}\right]}{\mathbb{E}\left[\lambda_i\right]} - \mathbb{E}\left[\mathcal{E}_{i,cross}\right] + \frac{N}{(N-1)^2}\mathbb{E}\left[\left(\frac{\lambda_i}{\mathbb{E}\left[\lambda_i\right]} - 1\right)\mathcal{E}_{i,cross}\right] \tag{92}$$

We can rewrite the middle term of equation (92) as:

$$\mathbb{E}_i\left[\mathcal{E}_{i,cross}\right] = \mathbb{E}_i\left[\frac{\mathbb{E}_{j\neq i}\left[\lambda_j \mathcal{E}_{ij}\right]}{\mathbb{E}_{j\neq i}\left[\lambda_j\right]}\right] \tag{93}$$

$$= \mathbb{E}_{j\neq i}\left[\frac{\mathbb{E}_{i}\left[\lambda_j \mathcal{E}_{ij}\right]}{\mathbb{E}_{j\neq i}\left[\lambda_j\right]}\right] \tag{94}$$

$$= \mathbb{E}_j\left[\frac{\lambda_j}{\mathbb{E}_j\left[\lambda_j\right]}\mathbb{E}_{i\neq j}\left[\frac{\mathbb{E}_j\left[\lambda_j\right]}{\mathbb{E}_{j\neq i}\left[\lambda_j\right]}\mathcal{E}_{ij}\right]\right] \tag{95}$$

$$= \mathbb{E}_j\left[\frac{\lambda_j}{\mathbb{E}_j\left[\lambda_j\right]}\bar{\mathcal{E}}_{.j}\right] \tag{96}$$

The term $\bar{\mathcal{E}}_{.j}$ resembles an equal-weighted average over rows of $\mathcal{E}_{ij}$ excluding the diagonal element.

The right term of equation (92) is a small-sample term that goes to zero in N:

$$\operatorname*{plim}_{N\to\infty} \widehat{\mathcal{E}} = \mathbb{E}\left[\omega_i\left(\mathcal{E}_{ii} - \bar{\mathcal{E}}_{.i}\right)\right] \tag{97}$$

$$\omega_i = \frac{\lambda_i}{\mathbb{E}\left[\lambda_i\right]} \tag{98}$$

is a λ_i weighted average of own- and cross elasticities.

$$\tag{99}$$

A.2 Proof of Proposition 1

Start from the general demand equation with demand shocks:

$$\Delta D_i = \mathcal{E}_{ii}\Delta P_i + \sum_{j\neq i}\mathcal{E}_{ij}\Delta P_j + \Delta\bar{D}_i \tag{100}$$

- **Assumption 1.** $X_i = X_j \Rightarrow \mathcal{E}_{il} = \mathcal{E}_{jl} = \mathcal{E}_{\text{cross}}(X_i, X_l) = X_i'\mathcal{E}_S X_l, \quad \forall i,j \in \mathcal{S}, l \neq i,j,$ where X_i is a $K \times 1$ vector of observables, and $\mathcal{E}_S$ is a $K \times K$ matrix.
- **Assumption 2.** $\mathcal{E}_{ii} - \mathcal{E}_{\text{cross}}(X_i, X_i) = \mathcal{E}_{jj} - \mathcal{E}_{\text{cross}}(X_j, X_j) = \widehat{\mathcal{E}}, \quad \forall i,j \in \mathcal{S}$

Proposition 7 *Under Assumption 1 and 2 and the exogeneity condition, the IV estimator, conditioning on X_i, identifies coefficient $\widehat{\mathcal{E}}$.*

Proof. Starting from equation (100), we can rewrite the demand equation as a cross-sectional regression:

$$\begin{aligned}
\Delta D_i &= \mathcal{E}_{ii}\Delta P_i + \sum_{j\neq i}\mathcal{E}_{ij}\Delta P_j + \Delta\bar{D}_i && (101)\\
&= \mathcal{E}_{ii}\Delta P_i + \sum_{j\neq i}\mathcal{E}_{cross}(X_i, X_j)\Delta P_j + \Delta\bar{D}_i && (102)\\
&= \left(\mathcal{E}_{ii} - \mathcal{E}_{cross}(X_i, X_i)\right)\Delta P_i + \sum_{j}\mathcal{E}_{cross}(X_i, X_j)\Delta P_j + \Delta\bar{D}_i && (103)\\
&= \widehat{\mathcal{E}}\Delta P_i + \sum_{j}\mathcal{E}_{cross}(X_i, X_j)\Delta P_j + \Delta\bar{D}_i && (104)\\
&= \widehat{\mathcal{E}}\Delta P_i + \sum_{j}X_i'\mathcal{E}_S X_j\Delta P_j + \Delta\bar{D}_i && (105)\\
&= \widehat{\mathcal{E}}\Delta P_i + \underbrace{X_i'\left(\sum_{j}\mathcal{E}_S X_j\Delta P_j\right)}_{\equiv\theta} + \Delta\bar{D}_i && (106)\\
&= \widehat{\mathcal{E}}\Delta P_i + \theta' X_i + \Delta\bar{D}_i && (107)
\end{aligned}$$

Equation (103) adds and subtracts $\mathcal{E}_{cross}(X_i, X_i)\Delta P_i$. Equations (104) and (105) use assumptions 2 and 1, respectively. Equation (106) pulls out X_i' from the sum. The remaining part of the sum gets absorbed into θ, a $K \times 1$ vector of cross-sectional constants. These θ are K regression coefficients on the K observables, X_{ik}.

Given the exclusion restriction that $Z_i \perp \Delta\bar{D}_i | X_i$ and the relevance condition that $cov(\Delta P_i, Z_i | X_i) \neq 0$, this is the standard IV setting, and the regression estimates $\widehat{\mathcal{E}}$. ■

Proposition 8 *Assumptions 1 and 2 are equivalent to an elasticity matrix $\mathcal{E}$ with:*

$$\mathcal{E} = \widehat{\mathcal{E}}\mathbf{I} + X\mathcal{E}_S X' \tag{108}$$

Proof. Write out matrix $\mathcal{E}$:

$$\mathcal{E} = \begin{pmatrix} \mathcal{E}_{11} & \mathcal{E}_{12} & \dots & \mathcal{E}_{1N} \\ \mathcal{E}_{21} & \mathcal{E}_{22} & \dots & \mathcal{E}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{E}_{N1} & \mathcal{E}_{N2} & \dots & \mathcal{E}_{NN} \end{pmatrix} = \begin{pmatrix} \widehat{\mathcal{E}} + X_1'\mathcal{E}_S X_1 & X_1'\mathcal{E}_S X_2 & \dots & X_1'\mathcal{E}_S X_N \\ X_2'\mathcal{E}_S X_1 & \widehat{\mathcal{E}} + X_2'\mathcal{E}_S X_2 & \dots & X_2'\mathcal{E}_S X_N \\ \vdots & \vdots & \ddots & \vdots \\ X_N'\mathcal{E}_S X_1 & X_N'\mathcal{E}_S X_2 & \dots & \widehat{\mathcal{E}} + X_N'\mathcal{E}_S X_N \end{pmatrix} \quad (109)$$

The (i,j) element of matrix $\mathcal{E}$ is $[\mathcal{E}]_{ij} = X_i'\mathcal{E}_S X_j = \mathcal{E}_{cross}(X_i, X_j)$, as defined by Assumption 1, for $i \neq j$. The diagonal elements are $[\mathcal{E}]_{ii} = \widehat{\mathcal{E}} + X_i'\mathcal{E}_S X_i = \widehat{\mathcal{E}} + \mathcal{E}_{cross}(X_i, X_i)$, as in Assumption 2. Since each element in $\mathcal{E}$ directly corresponds to the respective $\mathcal{E}_{ij}$ defined by Assumptions 1 and 2, the assumptions are equivalent to the elasticity matrix in (108). ■

Proposition 9 *Under Assumption 1 and 2, the multipler matrix $\mathcal{M} = -\mathcal{E}^{-1}$ also satisfies assumptions* 1 *and* 2*, with $\widehat{\mathcal{M}} = -1/\widehat{\mathcal{E}}$.*

Proof. Starting from equation (108), and applying the Woodburry matrix identity:

$$-\mathcal{E}^{-1} = -\left(\widehat{\mathcal{E}}\mathbf{I} + X\mathcal{E}_S X'\right)^{-1} \quad (110)$$

$$= -\widehat{\mathcal{E}}^{-1}\mathbf{I} + X\left(\widehat{\mathcal{E}}^2\mathcal{E}_S^{-1} + \widehat{\mathcal{E}}X'X\right)^{-1}X' \quad (111)$$

$$= \widehat{\mathcal{M}}\mathbf{I} + X\mathcal{M}_S X' \quad (112)$$

This corresponds exactly to assumption 1 and 2 applied to $\mathcal{M}$ with $\widehat{\mathcal{M}} = -1/\widehat{\mathcal{E}}$. ■

A.3 Proof of Proposition 5

Let $\mathcal{M}$ be defined as:

$$\mathcal{M} = \widehat{\mathcal{M}}\mathbf{I} + \left[\frac{1}{\mathbf{1}'\mathbf{1}}\mathbf{1}, X\right]\mathcal{M}_S\left[\frac{1}{\mathbf{1}'\mathbf{1}}\mathbf{1}, X\right]' \quad (113)$$

$$= \widehat{\mathcal{M}}\mathbf{I} + \frac{1}{(\mathbf{1}'\mathbf{1})^2}\mathcal{M}_{S[1,1]}\mathbf{1}\mathbf{1}' + \frac{1}{\mathbf{1}'\mathbf{1}}\mathcal{M}_{S,[1,2]}\mathbf{1}X' + \frac{1}{\mathbf{1}'\mathbf{1}}\mathcal{M}_{S,[2,1]}X\mathbf{1}' + \mathcal{M}_{S,[2,2]}XX' \quad (114)$$

where X is a $N \times 1$ vector of just one observable that satisfies $\mathbf{1}'X = 0$ and $X'X = 1$, and where $\mathcal{M}_S$ is a 2×2 matrix.

Furthermore, let

$$\Delta P = \mathcal{M}\Delta D \quad (115)$$

$$\Delta D = Z + u, \quad (116)$$

where Z is a $N \times 1$ vector of exogeneous instruments and u is a $N \times 1$ vector of unobserved shifts of the demand curve, and let $Z \perp \{u | X\mathbf{1}_t, \mathbf{1}_t\}$.

After substituting:

$$
\begin{aligned}
\Delta\tilde{P} &\equiv \Delta P - \widehat{M}Z && (117)\\
&= \mathcal{M}\Delta D - \widehat{M}Z && (118)\\
&= \widehat{\mathcal{M}}u + \frac{1}{(\mathbf{1}'\mathbf{1})^2}\mathcal{M}_{S[1,1]}\mathbf{1}\mathbf{1}'\Delta D + \frac{1}{\mathbf{1}'\mathbf{1}}\mathcal{M}_{S,[1,2]}\mathbf{1}X'\Delta D + \frac{1}{\mathbf{1}'\mathbf{1}}\mathcal{M}_{S,[2,1]}X\mathbf{1}'\Delta D + \mathcal{M}_{S,[2,2]}XX'\Delta D && (119)\\
&= \widehat{\mathcal{M}}u + \mathcal{M}_{S[1,1]}\frac{\mathbf{1}}{\mathbf{1}'\mathbf{1}}\left(Z_{agg} + u_{agg}\right) + \mathcal{M}_{S,[1,2]}\frac{\mathbf{1}}{\mathbf{1}'\mathbf{1}}\left(Z_X + u_X\right) + \mathcal{M}_{S,[2,1]}X\left(Z_{agg} + u_{agg}\right) + \mathcal{M}_{S,[2,2]}X\left(Z_X + \right. && (120)
\end{aligned}
$$

where $Z_{agg} = \frac{\mathbf{1}'Z}{\mathbf{1}'\mathbf{1}}$, $Z_X = X'Z$, $u_{agg} = \frac{\mathbf{1}'u}{\mathbf{1}'\mathbf{1}}$, and $u_X = X'u$. Further let $\Delta\tilde{P}_{agg} = \mathbf{1}'\Delta\tilde{P}$ and $\Delta\tilde{P}_X = X'\Delta\tilde{P}$, such that:

$$
\begin{aligned}
\Delta\tilde{P}_{agg} &= \widehat{\mathcal{M}}\mathbf{1}'\mathbf{1}u_{agg} + \mathcal{M}_{S[1,1]}(Z_{agg} + u_{agg}) + \mathcal{M}_{S,[1,2]}(Z_X + u_X) && (121)\\
&= \mathcal{M}_{S[1,1]}Z_{agg} + \mathcal{M}_{S,[1,2]}Z_X + \left(\widehat{\mathcal{M}}\mathbf{1}'\mathbf{1} + \mathcal{M}_{S[1,1]}\right)u_{agg} + \mathcal{M}_{S,[1,2]}u_X && (122)\\
\Delta\tilde{P}_X &= \widehat{\mathcal{M}}u_X + \mathcal{M}_{S,[2,1]}(Z_{agg} + u_{agg}) + \mathcal{M}_{S,[2,2]}(Z_X + u_X) && (123)\\
&= \mathcal{M}_{S,[2,1]}Z_{agg} + \mathcal{M}_{S,[2,2]}Z_X + \mathcal{M}_{S,[2,1]}u_{agg} + \left(\mathcal{M}_{S,[2,2]} + \widehat{\mathcal{M}}\right)u_X && (124)
\end{aligned}
$$

For $Z_{agg}, Z_X \perp u_{agg}, u_X$, this corresponds to two time-series regressions of $\Delta\tilde{P}_{agg}$ and $\Delta\tilde{P}_X$ on Z_{agg} and Z_X. These regressions identify all elements of $\mathcal{M}_S$.

A.4 Proof of Proposition 4

The proof is constructive and follows naturally from aggregating the demand shock together.

We consider a matrix $\mathcal{M}$ which satisfies assumption A1. We note $\mathcal{M}_{ii} = \mathcal{M}_{own,i}$. Then all other coefficients are obtained by combining the own elasticity with the relative elasticity: $\mathcal{M}_{ji} = \mathcal{M}_{own,i} - \widehat{\mathcal{M}}$ for all $j \neq i$.

Choose a vector of weights θ such that $\theta'\mathbf{1} = 1$. We can generically decompose changes in demands as:

$$
\begin{aligned}
&\Delta D = \Delta D_{\text{agg}} \cdot \mathbf{1} + \Delta D_{\text{idio}}, \quad \text{with } \theta'\Delta D_{\text{idio}} = 0 && (125)\\
\iff\quad &\Delta D_{\text{agg}} = \theta'\Delta D && (126)
\end{aligned}
$$

Let us multiply each component by $\mathcal{E}$:

$$\mathcal{M}\mathbf{1} = \underbrace{\left(\sum_i \mathcal{M}_{\text{own},i} - (N-1)\widehat{\mathcal{M}}\right)}_{\bar{\mathcal{M}}}\mathbf{1} \tag{127}$$

$$[\mathcal{M}\mathit{\Delta} D_{\text{idio}}]_i = \sum_j \mathcal{M}_{ij}\mathit{\Delta} D_{\text{idio},j} = \mathcal{M}_{\text{own},i}\mathit{\Delta} D_{\text{idio},i} + \sum_{j\neq i}\left(\mathcal{M}_{\text{own},j} - \hat{\mathcal{M}}\right)\mathit{\Delta} D_{\text{idio},j} \tag{128}$$

$$= \widehat{\mathcal{M}}\mathit{\Delta} D_{\text{idio},i} + \sum_j (\mathcal{M}_{\text{own},j} - \widehat{\mathcal{M}})\mathit{\Delta} D_{\text{idio},j} \tag{129}$$

$$= \widehat{\mathcal{M}}\mathit{\Delta} D_{\text{idio},i} + (\mathcal{M}_{\text{own}} - \widehat{\mathcal{M}}\mathbf{1})'\mathit{\Delta} D_{\text{idio}} \tag{130}$$

The second term is constant across index is such that when we rewrite it in vector form:

$$\mathcal{M}\mathit{\Delta} D_{\text{idio}} = \widehat{\mathcal{M}}\mathit{\Delta} D_{\text{idio}} + \mathbf{1}(\underbrace{\mathcal{M}_{\text{own}}}_{\text{vector}} - \widehat{\mathcal{M}}\mathbf{1})'\mathit{\Delta} D_{\text{idio}} \tag{131}$$

$$\mathcal{M}\mathit{\Delta} D_{\text{idio}} = \left(\widehat{\mathcal{M}}\mathbf{I} + \mathbf{1}(\mathcal{M}_{\text{own}} - \widehat{\mathcal{M}}\mathbf{1})'\right)\mathit{\Delta} D_{\text{idio}} \tag{132}$$

We see that if θ is proportional to $(\mathcal{M}_{own} - \widehat{\mathcal{M}}\mathbf{1})$, the second term disappears, which completes the proof.

A.5 Proof of Proposition 6

We consider the case of G groups. We define the matrix $\mathcal{E}_{gh}$ of size $N_g \times N_h$

$$\mathcal{E}_{gh} = [\mathcal{E}_{ij}]_{\{i\in g, j\in h\}} \tag{133}$$

As a consequence the matrix $\mathcal{E}_{gg}$ is the block for subgroup g.

$$\begin{bmatrix} \mathcal{E}_{gg,g=1} & \mathcal{E}_{gh,g=1,h=2} & \cdots & \mathcal{E}_{gh,g=1,h=G} \\ \mathcal{E}_{gh,g=2,h=1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathcal{E}_{gh,g=1,h=G-1} \\ \mathcal{E}_{gh,g=G,h=1} & \cdots & \mathcal{E}_{gh,g=G-1,h=1} & \mathcal{E}_{gg,g=G} \end{bmatrix}$$

In the previous section we show how to deal with aggregation of the diagonal block, i.e. the one asset case. Now let us examine the off-diagonal blocks: $\mathcal{E}_{gh}$ for $g \neq h$. From A2, we have $\mathcal{E}_{ik} = \mathcal{E}_{jk}$, for $i, j \in g$ and $k \in h$. This means that $\mathcal{E}_{gh}$ has constant columns and can be written as:

$$\mathcal{E}_{gh} = \mathbf{1}\theta'_{gh} \tag{134}$$

This equation defines the vector of size N_h, θ_{gh}

$$\varDelta D_g = \mathcal{E}_{gg}\varDelta P_g + \sum_{h\neq g}\mathcal{E}_{gh}\varDelta P_h \tag{135}$$

$$\varDelta D_g = \underbrace{\hat{\mathcal{E}}_g\varDelta P_{\text{idio},g} + \bar{\mathcal{E}}_{gg}\underbrace{\theta'_{gg}\varDelta P_g}_{P_{\text{agg},g}}\mathbf{1}}_{\text{from prop. 3}} + \sum_{h\neq g}\mathbf{1}\theta'_{gh}\varDelta P_h \tag{136}$$

We use the θ_{gg} from the one group case of the section above to get the aggregate price effect $P_{\text{agg},g}$. We can project θ_{gh} on θ_{hh} (of size N_h):

$$\theta_{gh} = \bar{\mathcal{E}}_{gh}\theta_{hh} + \xi_{gh}, \quad \text{with } \theta'_{hh}\xi_{gh} = 0 \tag{137}$$

$$\iff \bar{\mathcal{E}}_{gh} = \left(\theta'_{hh}\theta_{hh}\right)^{-1}\theta'_{hh}\theta_{gh} \quad \text{(regression formula)} \tag{138}$$

The formula simplifies to:

$$\varDelta D_g = \hat{\mathcal{E}}\varDelta P_{\text{idio},g} + \bar{\mathcal{E}}_{gg}\varDelta P_{\text{agg},g} + \sum_{h\neq g}\bar{\mathcal{E}}_{gh}\theta'_{hh}\varDelta P_h\mathbf{1} + \sum_{h\neq g}\mathbf{1}\xi'_{gh}\varDelta P_h \tag{139}$$

$$\varDelta D_g = \hat{\mathcal{E}}\varDelta P_{\text{idio},g} + \bar{\mathcal{E}}_{gg}\varDelta P_{\text{agg},g} + \sum_{h\neq g}\bar{\mathcal{E}}_{gh}\varDelta P_{\text{agg},h}\mathbf{1} + \sum_{h\neq g}\mathbf{1}\xi'_{gh}\varDelta P_h. \tag{140}$$

If we assume that $\xi_{gh} = 0$, that is θ_{gh} is proportional to θ_{hh} (both vectors of size N_h), we have a well-defined aggregated system:

$$\varDelta D_{\text{agg}} = \bar{\mathcal{E}}\varDelta P_{\text{agg}}, \tag{141}$$

with

- $\bar{\mathcal{E}}_{gg}$ and θ_{gg} are defined by applying proposition 3 to the block $\mathcal{E}_{gg}$ (group g).
- The off-diagonal elements $\bar{\mathcal{E}}_{gh}$ are given by applying equation (137) given the definition of θ_{gh} in (134)

The spillovers from the asset in group h on the asset in group g is the same for all assets in group g. It is equal to

$$\bar{\mathcal{E}}_{gh}\bar{P}_h + \vartheta'_{gh}\mathbf{P}_{\text{individual},h}, \qquad \text{with } \mathbf{1}'_h\mathbf{P}_{\text{individual},h} = 0 \tag{142}$$

In general if $\vartheta_{gh} \neq 0$, both the average price change in group h and the distribution of the price changes within assets of group h determine the spillover (of h on g).

If we further impose that $\vartheta_{gh} = 0$, the second effect disappears and the spillover is determined by $\bar{\mathcal{E}}_{gh}$ (the sum of the rows of the off-diagonal block) $\vartheta_{gh} = 0$ can be a consequence of imposing a form of symmetry in the off-diagonal blocks; either through having the elasticity matrix be symmetric joint with the group version of assumption **A2**. We can also get $\vartheta_{gh} = 0$ from group A2 and $\mathcal{E}_{gh} = \mathcal{E}'_{hg}$ (this is weaker than full symmetry because this does not require the diagonal blocks themselves to be diagonal).