Title
Essays on Asset Pricing and Financial Institutions

Author
Kiefer, Patrick Christian

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Essays on Asset Pricing and Financial Institutions

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy in Management

by

Patrick Christian Kiefer

2018

ABSTRACT OF THE DISSERTATION

Essays on Asset Pricing and Financial Institutions

Patrick Christian Kiefer

Doctor of Philosophy in Management

University of California, Los Angeles, 2018

Professor Mark S. Grinblatt, Chair

Chapter 1 Abstract. Forecasts of risk prices at alternative time scales can be used to consolidate history dependence in asset return time series. The resulting Markovian structure identifies a martingale component in the latent transition dynamics. I apply the model to U.S. stock markets and find the concentration of return volatility on the martingale component - *the spectral gap* - is countercyclical, and predicts annual market returns out-of-sample (o.o.s.) with an R^2 of 10.8%. Value (HML) predictability is concave and front-heavy, peaking at a one-year 14.7% o.o.s. R^2. In contrast, the momentum predictability term structure is convex, insignificant on the short end, but accelerates to 31.4% o.o.s. R^2 at the three-year horizon. I form *timing portfolios* to investigate the risk content of the aggregate forecasts. Incremental gains from timing value are compensation for bearing systematic shocks to time-varying expected returns. Exposure to the market timing portfolio is cross-sectionally priced, while gains from timing size (SMB) are not. The findings provide new restrictions for parametric asset pricing theories.

Chapter 2 Abstract. Incomplete human capital markets induce unexpected rebalancing costs that are mitigated by a bank. Ex-ante, the bank exchanges risky endowments for demandable liabilities. An ex-post withdrawal corresponds to exercising a put option on the market, used to resolve an unexpected portfolio choice problem. Portfolio choice opens a risk aversion channel that distinguishes our predictions from Diamond and Dybvig (1983) and related models. In these models, deposits resolve consumption-timing tensions by accommodating the investor's intertemporal elasticity of substitution (IES). The inclusion of risk-based incentives allow us to characterize the endogenous link between the intermediary balance sheet and the preference-based pricing kernel. Moreover, ex-post rebalancing incentives relax enforcement problems for ex-ante optimal policies in incomplete markets. This provides a justification for the coexistence of intermediation and market institutions.

The dissertation of Patrick Christian Kiefer is approved.

Daniel Dumitru Andrei

Andrea Lynn Eisfeldt

Pierre-Olivier Weill

Mark S. Grinblatt, Committee Chair

University of California, Los Angeles

2018

Dedication

To Ariana

Contents

List of Figures

C.Vita

Patrick Christian Kiefer

Prior Education and Activities

2011, M.A., Economics, University of California, Los Angeles

2011, B.A., Economics, Honors, *Summa Cum Laude*, University of California, Los Angeles

2011, B.S., Applied Mathematics, University of California, Los Angeles

2007-2008 City College of San Francisco

San Francisco Fire Department, Potrero Hill Emergency Response Team, 2005-2007

Founder and Principal, M.I. Incorporated, 2003-2007

Chapter 1

The Factor Structure of Time-Series Predictability

1.1 Introduction

Linear risk factor models reduce large complex markets to a handful of portfolios, making empirical and theoretical studies of asset markets tractable and efficient. Dimension reduction is a structural implication of equilibrium asset pricing theory - the so-called *factor replicating portfolios* are sufficient because the distribution of any security's returns can be characterized in terms of its exposure to these portfolios. While this simplification is exploited routinely in the context of unconditional factor modeling, it is rarely used to reduce complexity in problems of predictability.

This is not because equilibrium asset pricing theory is silent on reducing dimension in the time series. A well known implication is easily seen in a constant relative risk aversion (CRRA) representative agent economy with stochastic volatility. In this model, the conditional volatility of the pricing kernel is simply the risk aversion coefficient times the volatility of aggregate consumption growth. Predictable variation of any asset's excess return is given by its beta times the conditional volatility of the pricing kernel. The dimension reduction in this model's forecasting problem underscores a general implication. If the conditional means

of the factor replicating portfolios vary through time, forecasts of these means are sufficient to forecast the excess returns of arbitrary portfolios.

We exploit these implications to produce a parsimonious method for characterizing conditional mean returns in real time. The method is simple to implement using principal components analysis (PCA) and single-lag vector autoregressions VAR(1), making it accessible to practitioners and regulators. By appealing to the pricing kernel structure in the time series, we exploit the theoretical implication linking the dynamics of aggregate risk to the systematic components of predictability. Several positive implications related to the method's success are presented.

The primary challenge to implementing such a model empirically is the estimation of conditional means. We address this by exploiting a first-order transition representation for the underlying dynamics. A key ingredient is the information set for realized returns that collapses higher-order history-dependence into a Markov state vector. This information set includes projections of historical returns. From here, a decomposition of the transition dynamics separates martingale shocks from transitory variation. The common sources of transitory variation identify the factor structure of time-varying expected returns.

We show that a simple statistic constructed from the latent expected return factors - the *spectral gap* - predicts the market with an out-of-sample R^2 of 10.8% annually. The spectral gap is also used to predict portfolio returns. We forecast the value (HML) and momentum (UMD) premia with annual o.o.s. R^2 's of 14.7% and 6.7% respectively. The forecasts highlight cross-sectional differences in predictability term structures. At the two -year horizon, momentum predictability reaches 14.5%. While market and momentum forecasts exhibit increasing rates of explained variation, the value premium is most predictable at a 1-year horizon.

We investigate the theoretical implications of our findings in two ways. First, for each of the Fama and French risk factors, we replicate the predictive series obtained from the spectral gap within the space of test assets. This allows us to quantify the efficiency gains relative

to the underlying factor models. Like the forecasts, the *timing portfolios* are constructed in real time. Unlike the forecasts, the timing portfolios are tradeable. We find returns to the timing portfolios have higher ex-ante Sharpe ratios and higher average realized Sharpe ratios than the underlying factors.

Second, we construct a time series of conditionally unpredictable returns from each of the timing portfolios. We test whether exposures to these shocks can explain cross-sectional variation in returns. In particular, using the underlying factor replicating portfolios as benchmarks, we isolate incremental changes to cross-sectional performance. Gains in efficiency from the market, momentum and value timing portfolios are priced in the cross-section, while gains from size (SMB) timing portfolios are not.

Classically, an empirical risk factor is formed in two steps. Traded assets are sorted into bins according to a characteristic, such as book-to-market equity. Then, a zero-cost portfolio is formed by shorting the lowest bin and buying the highest bin. When the sorting characteristic proxies for exposure to an undiversifiable source of risk, mean returns to the zero-cost portfolio are proportional to the unconditional price of exposure to this risk. The resulting *factor replicating portfolios* form the basis of empirical risk factor models.

However, conditional and unconditional expected returns differ systematically. The dividend yield, *cay*, and other variables forecast market returns, producing estimates of the time-varying expected returns.[1] In equilibrium the expected excess return on the market is equal to the market risk price squared (uniquely up to changes in basis). Similarly, portfolio-level forecasts can be used to estimate time-varying risk prices for value (HML), size (SMB), momentum (UMD) and other non-market factors. Importantly, systematic variation in expected returns contains the time-variation in factor risk prices. When all systematic variation is priced, the two are equivalent.

Evidence suggests the term structure of equity risk is not flat, and that the sign of the

[1]The series *cay* is constructed in Lettau and Ludvigson (2001) to capture deviations in the consumption-wealth ratio from its long-run mean.

slope changes depending on economic conditions.[2] Information in the full term structure may be relevant for forecasting single-period returns, but the exact term structure is not observed and no consensus estimation method exists. To capture this information while remaining agnostic about the structural model, we construct return forecasts of priced risk factors for several horizons separately. Henceforth we refer to these forecasts as *nominal* forecasts.

To estimate time-varying risk prices, we augment asset returns with nominal forecast errors, fitted within each rolling window separately for several horizons. The nominal forecasts we use are fitted values of Fama-French risk-factor returns on standard (lagged) predictors, such as the dividend yield and past returns. We reject two important null hypotheses that characterize nominal forecasts. The first is that nominal forecasts over different horizons are deterministic functions of each other. The second is that every nominal forecast is replicable in the space of (contemporaneous) test assets.

The signals in the nominal forecast errors are a key input for extracting the latent factors and risk price estimates. Consider an event of the sort: "nominal market forecast errors at the 4-year horizon tend to be high when one-month HML returns are low." The information in this event is *not* contained in contemporaneous realized returns if the sequence of nominal 4-year market forecast errors *cannot* be replicated by a portfolio of test assets.[3] Conditioning on these events is valuable empirically. In addition to refining the information set, nominal forecast errors are chosen in exactly the linear combinations of past returns that justify a Markov representation.

Given a Markov representation of returns - including past returns, in the case of U.S. equities - we run a classical principal components analysis for each period over a fixed history length. The PCA maps to a decomposition of a generic vector Euler equation when dynamics

[2]Binsbergen and Koijen (2016) provide a thorough survey.

[3]Specifically, today's realization of the nominal 4-year market forecast error is the difference between today's one-period realized market return and the prediction of that return made 4-years ago.

are Markovian.[4] The history length is chosen to isolate fluctuations at a particular distance from statistical equilibrium, controlled by the so-called *mixing times*. The representation is updated in each period by repeating this construction.

The construction captures changes in the risk price and composition of transitory factors while keeping the scale of the *average* transitory fluctuation fixed at a constant fraction of total volatility. Technically, for a given threshold, we keep the mean mixing time fixed. Changes in the risk price and composition of transitory factors correspond to *systematic fluctuations in expected returns*. We find that while an average of 86% of quarterly return volatility is concentrated on permanent shocks to market capitalization, the forecasting power comes from incorporating the systematic predictable variation in expected returns - *the expected return factors*.

Writing the Euler equation in terms of decomposed returns shows the *spectral gap* is informative about future expected returns. A naive test of predictability using the lagged spectral gap predicts quarterly, semi-annual and annual market returns with out-of-sample R^2s of 3.8%, 6.5% and 10.8%. More sophisticated statistics and existing predictors are no better than the lagged spectral gap out of sample, with *cay* coming in second with at an annual o.o.s. R^2 of 7.6%. Gap statistics also forecast portfolio returns, predicting value (HML) with an annual out of sample R^2 of 14.7% and Momentum (UMD) with an R^2 of 6.4% (16.1% biannually).

Portfolio-level predictability term structures are significantly heterogeneous. Explained variation increases monotonically with horizon for the market and momentum risk factors. For value, explained variation is hump-shaped, with a maximum R^2 corresponding to the one-year forecasting horizon and decreasing afterward. We find the market loads with a coefficient 0.98 on shocks to the leading component. The leading component's autoregressive coefficient is indistinguishable from zero (the point estimate is 0.089 with s.e. 0.072). In contrast, value returns load significantly on the penultimate factor. The penultimate factor

[4] A formal description of this procedure is given in section 5.5.1.

tracks common variation in expected returns - the autoregressive coefficient is 0.649 with an adjusted standard error of 0.092.

In addition to the concave predictability term structure for the value premium, we find the momentum term structure is increasing and convex. Forecast horizons inside of 1 year feature rapidly increasing predictability in the value premium, while momentum is almost unpredictable. Forecast horizons longer than one year feature decreasing predictability of the value premium and simultaneous rapid increases in predictability of the momentum premium. Predictability of market expected returns increase linearly over the same forecasting horizons.

Value-weighted dividend yields play a key role in the construction of productive nominal forecasts. Using the value weighted CRSP index ex-dividend in place of dividend yields when constructing the nominal forecasts significantly restricts the predictive power for each factor other than size. However, predictability for the market picks up at the long end, suggesting capital gains have a small but significant role in positively predicting aggregate expected returns over lower frequencies. The equal-weighted CRSP ex-dividend index input generates predictive power for size but insignificant predictive power for the value premium, and significantly reduced predictive power for the market and momentum premiums.

Our findings are consistent with existing evidence on predictability and dividend yields. Nominal forecasts using single lags of dividend yields produce factor return forecasts that are inferior but nonetheless dominate dividend yield forecasting directly. Dividend yields are highly persistent, fluctuating with a half-life of roughly 15 years, and forecast returns. From this standpoint, that the full term structure of forecast errors is informative about expected returns is unsurprising - it is certainly informative about future yields.

We find the spectral gap is a meaningful macroeconomic indicator on its own. Dynamics of the spectral gap measure changes in the concentration of volatility on the leading priced risk factor. In the US stock markets, we find return volatility concentrates countercyclically on permanent, or "long-run" shocks. These findings have implications for parametric

stochastic discount factor (SDF) models. Because the spectral gap is an accessible object in any Markov model of asset prices, these findings can help discriminate among competing parametric theories.

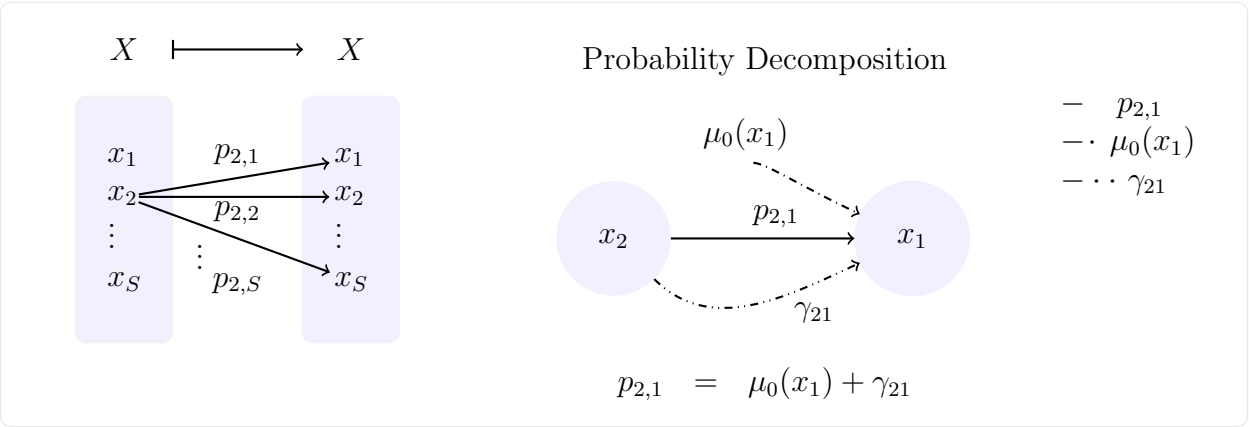
The prices of incremental gains from timing portfolios are informative about the risk content of the underlying factor. By projecting the factor forecasts back on to the space of test assets, we limit our analysis to changes in the distribution of risk across portfolios because the total risk is constrained to be the value-weighted excess returns of the test assets. Relative efficiency gains measure the *intensive* margin of factor risk. Marginal efficiency gains arise when marginal and average risk prices diverge. We reject that the marginal prices of exposure to the value and momentum portfolio returns are zero. We cannot reject that the marginal price of exposure to size is zero.

The study proceeds with a discussion of the literature, followed by a Markov model of asset returns. Section 4 describes the empirical tests derived from the Markov model. The data and empirical results are reported in sections 4.3 and 4.4. Appendices contain technical details and a handful of ancillary charts and tables.

1.2 Related Work

Alvarez and Jermann (AJ) (2005), and Hansen and Scheinkman (2011) factorize the pricing kernel into martingale and transitory components. (AJ) find that volatility of the growth rate of the martingale is roughly 90% of the volatility of the stochastic discount factor. Hansen and Scheinkman (2011) use the Perron-Frobenius theory to isolate the asymptotic risk-return tradeoff for an aggregate payoff functional when the underlying dynamics are Markovian. Borovicka, Hansen and Scheinkman (2016) point out the structure and interpretation of the Perron-Frobenius estimate depends heavily on model specification. We use a Perron-Frobenius decomposition to identify predictable fluctuations in latent factors that are

Figure 1.1: Decomposition over Finite State Space



(a) Predictable fluctuations are identified using a decomposition of transition probabilities. The probability of moving to x_1 given the current state x_2 is $p_{2,1} = \mu_0(x_1) + \gamma_{21}$. The invariant probability $\mu_0(x_1)$ does not depend on x_2, and dominates forecasts asymptotically. The transitory correction γ_{21} becomes negligible in large time, but contributes importantly to local dynamics. Perron-Frobenius and spectral theories provide decompositions of the signed measures $\nu : X \times X \mapsto \mathbb{R}$ rather than the state space X itself. This point is clear for finite-state ergodic processes, pictured here, where it does not make sense to claim subsets of finitely many points are transitory: all points in an ergodic set are visited with probability one in large time.

negligible asymptotically. These components constitute roughly 15% of the total variation in returns, corroborating findings in (AJ).

Chen, Roll and Ross (1986) find that exposure to innovations in the term spread, credit spread, and industrial productivity help explain cross-sectional variation in average stock returns. Hansen and Jagannathan (1991), (AJ), Ross (2015), and Backus and Chernov (2008) argue that while important work is done using macroeconomic variables to understand asset prices, it is also the case that equilibrium prices reveal information about the macroeconomy. In particular, Hansen and Jagannathan (1991) study restrictions placed by observed prices on the mean and variance of the pricing kernel and argue the pricing kernel must be an order of magnitude more volatile than consumption growth to justify the observed Sharpe ratios. Backus and Chernov (2008) study restrictions on pricing kernel cumulants implied by observed prices and use evidence of higher moments to rule out symmetric dynamics. Ross (2015) argues for recovery of the underlying physical transition dynamics from returns. To

our knowledge, identification of persistent dynamics in the pricing kernel from the covariance matrix of returns is a novel contribution.

Koijen, Lustig and Van Nieuwerburgh (2017) propose a three-factor with a separate role for business-cycle fluctuations to explain average returns to stock portfolios sorted on book-to-market equity and maturity-sorted treasury portfolios. Cochrane and Piazzesi (CP) show a single bond factor constructed from a linear combination of forward rates predicts returns on bond portfolios of any maturity and forecasts returns in equity markets. The (CP) factor describes transitory fluctuations in the Koijen, Lustig and Van Nieuwerburgh economy; the market factor captures permanent shocks to cash flow levels, and the third factor updates inflation expectations. Our findings corroborate and extend their classification of factors. We show over 90% of the variation in the market and the momentum (UMD) factor is driven by innovations with no transitory content, while over 70% of the variation in value (HML) is explained by the transitory factors.

Fama and French (1992, 1993) and Asness (1994) document the importance of book-to-market equity for explaining cross-sectional variability in average returns and establish the value (HML) and size (SMB) factors to augment the single-factor (CAPM) model. Jegadeesh and Titman (1993) and Moskowitz and Grinblatt (1999) study momentum by asset and industry, respectively, and find a cross-sectional ranking of past winners and losers incrementally improves dynamic efficiency properties of the Fama-French three factor models. Berk (1995) points out that if size is the expected cash flow level, and two firms have identical "size" but one has lower market cap, it is because it has a higher discount rate. Mechanically, size inversely predicts average returns and will appear to be priced whenever a true factor is missing. We find incremental efficiency gains from timing size are not priced, while gains from timing the market, momentum and value are priced.

Several papers argue for improvements in the prevailing constructions of Fama-French factors. Gerokos and Lihnnainmaa (2012) argue that the HML factor returns can be decomposed into price-driven and book-driven elements, and that only the price-driven component

of the HML factor returns can explain cross-sectional variation in returns. Asness and Frazzini (2013) argue that HML contains about 20% momentum, and propose a construction of HML that isolates the "pure value" component. Lihnnainmaa (2015, 2016) finds accrual, investment, and profitability factor constructions that are preferred to the Fama French (2015) constructions for cross-sectional asset pricing. We find evidence the value premium is compensation for shocks to the persistence of expected returns. In contrast, we find momentum is compensation for i.i.d. shocks to realized returns.

Binsbergen, Brandt, and Koijen (2012) synthesize dividend strips at various maturities to analyze the term structure of equity risk. They find that short-run cash flows have higher average returns that the market, implying a downward-sloping term structure of equity premia. Ait-Sahalia, Karaman, and Mancini (2015) provide evidence that the sign of the slope is time-varying and procyclical. Schulz (2016) argues the downward sloping term structure becomes insignificant when tax rates specific to dividend income are considered. Weber (2016) sorts stocks based on a measure of cash flow duration and finds high-duration stocks earn roughly a one-percent premium monthly over low-duration stocks, providing term-structure evidence that does not rely on synthesizing dividend payments. This set of findings motivate us to study nominal forecasts of returns on equity portfolios separately at different horizons.

Bandi and Tamoni (2015) implement a time-scale decomposition of returns and consumption growth by projection on the Haar basis. The result is a representation of the time-series of returns as the sum of moving averages over J non-intersecting intervals of increasing scale 2^j $j \in \{0, 1, ..., J\}$. Severino (2014) shows existence of time-series decompositions based on the sequential application of an isometric operator. The decomposition splits a Hilbert space into an infinite direct sum of rank-one prediction-error subspaces and that sum's (possibly empty) orthogonal complement. The derivation generalizes the Wold representation. Our decomposition is most related methodologically to Bandi and Tamoni (2015) and Severino (2014). The functional basis in our decomposition is identified from a PCA of a generalized covariance matrix of realized returns.

Bansal and Yaron (2005), and Hansen, Heaton and Li (2008) propose and evaluate slow-moving latent growth factors as explanations for unconditional risk premia. Bansal, Kiku and Yaron (2010) sharply characterize the distinction between long run risk - captured by shocks to persistent levels driving growth, and business cycle risk - captured by shocks to persistent growth directly. Hansen and Sargent (2007, 2016) outline the similarities between long-run risk and the implications of robust control policies in financial markets. The authors show the long-run risk model is the model a robust Epstein-Zinn investor would appear to have referenced ex-post. We contribute to this discussion by quantifying slow moving changes in the concentration of return volatility on long-run shocks. Estimates of this quantity predict portfolio returns and contribute to priced factor risk, suggesting a new layer of tests from within parametric long-run-risk or ambiguity aversion models.

Jagannathan and Wang (JW) (1996) take a CAPM model with conditionally dependent parameters and condition down. It is well known that the unconditional version includes correction parameters capturing the correlation between time- varying exposures and time-varying risk prices. JW circumvent the problem of obtaining conditional estimates directly by using the $BAA - AAA$ credit spread as a proxy for the time-varying risk price. They provide GMM estimates that support the conditional CAPM over the implied constant parameter CAPM tested classically. In U.S. stocks, we provide evidence the latent pricing kernel contains the required forecasting variables itself. As a diagnostic, we find including the $BAA-AAA$ credit spread in the calculation of the covariance matrix generates forecasts with o.o.s. R^2 gains of 0% to 2% annually, depending on the portfolio and subsample.

Cochrane (2011) surveys the in-sample evidence suggesting that, across many asset classes, low yields (high prices) today predict low returns in the future - and not cash flow growth. The importance of jointly restricting cash flow and yield parameters in a vector-autoregression (VAR) for evaluating the predictability of discount rates is given by Cochrane (2008). Santa Clara (2015) finds evidence that dividend yields in fact predict cash flows the portfolio-level. Brennan and Taylor (2016) find aggregate and portfolio-level return pre-

dictability out of sample. The authors distinguish risk-based sources of predictability from potentially non-equilibrium present-value predictors by comparing variables obtained from the covariance matrix with those obtained from replicating portfolios for present values of cash flow news and discount rate shocks. We base aggregate and portfolio-level o.o.s. forecasts on a small number of common factors - also advocated by Cochrane (2011). To quantify the risk content of our predictors, we test whether the incremental efficiency gains, measured by the difference in returns between a forecast replicating portfolio and the underlying target portfolio, are priced in the cross-section.

Chen (2005) points out that out of sample (o.o.s.) tests of predictability are well suited for studying time series containing one or more discrete structural breaks. Chen finds that o.o.s. predictive statistics capture discrete structural breaks in Taiwanese savings rates that predict declines in investment rates that followed. Importantly, the author shows an in-sample vector-autoregression model misses the break and fails to to reject the null of no-predictability. We provide diagnostics in section (5.5) indicating our o.o.s. predictors react to structural breaks quickly relative to rolling beta and rolling naive PCA models.

Goyal and Welch (2008) provide evidence that predictors in the literature perform poorly out of sample, and that prediction quality in-sample is often confined to crisis periods. However, although low in absolute terms, Goyal and Welch (2003) and (2008) find the relative forecasting power of the dividend-price ratio is highest using out of sample predictions over the postwar sample ending in 1990. They argue the incremental efficacy of the forecasts arise through the ability of rolling beta estimates to pick up changes in the data generating process (DGP). Similarly to Chen (2005), the authors find a constant coefficient VAR model fails to generate a significantly non-zero R^2. Our predictive efficacy is not limited to crises, suggesting we capture important cases where the underlying model changes smoothly but manifests as a significantly distinct process along sufficiently non-overlapping subsamples.

Kelly and Pruiit (2013) implement a three-pass filter to forecast portfolio returns using a large cross-section of predictor variables. They report test statistics derived to account

for the fitting procedure, and measure out-of-sample predictive R^2 by fitting the model to data omitting the target period and predicting the target period. R^2s on annual market predictability reach 13%. We report comparable but lower annual market R^2 of 10.8% without using cross-sectional data and without using future data. Thus, our decomposition is informative in real time. However, the predictive regressions proposed by Kelly and Pruiit (2013) apply to predictive settings where a Markov structure may not be warranted, while our identification relies on this structure.

Bryzgolva (2014) advocates for traded (price-based) proxies for risk factors over macro factors for statistical reasons. A constrained LASSO-style regression penalizes candidate risk factors with poorly measured exposures. Traded factor returns have low levels of idiosyncratic noise, making exposures easier to measure and thus the LASSO procedure penalizes traded proxies less. This helps rationalize the incremental gain in statistical significance from the timing portfolio returns over the extracted latent process. Simultaneously, it underscores the improvements in significance and precision of the timing portfolios relative to the conventional factor replicating portfolios.

Shrinkage estimation procedures - such as the LASSO for the covariance matrix - are a potentially relevant exercise in our setting. We avoid this by first considering the factor returns only, rather than the rank-deficient cross-sections with many test assets. However, the inclusion of nominal forecasts in the generalized covariance matrix estimation introduces rank deficiency. Moreover, one may shrink towards a target horizon Sharpe ratio when the data matrix includes many horizon-specific forecasts. For the purposes of this paper, we manage rank deficiency and sparseness by exploiting the inner-unitary properties of the rotation matrices from a singular value decomposition. Cases of explicit shrinkage are delegated to future work.

We draw on tools from Markov processes, large deviations and random matrix theory. The modern theory of large deviations is due to Donsker and Varadhan (1975a, 1975b) and Gartner (1977), building on Cramer (1938) and Sanov (1957). Exponentially unlikely devi-

ations of Markov processes are described by the Perron-Frobenius theory (Varadhan (1983, 2008), Hansen (2011), Borovicka, Hansen and Scheinkman (2016)). Brownian dynamics for the spectra of random matrices were introduced by Dyson (1962a, 1962b). Erdos and Yao (2017) and Tao (2011) characterize spectral dynamics for sequences of random matrices and nest Dyson Brownian motion as a special case. Knowles, Yao and Yin (2014) provide asymptotics for outlying eigenvalues of covariance matrices when the parameter dimension grows proportionally to sample size.

Interest in Markov-Chain Monte-Carlo (MCMC) methods drove a better understanding of convergence rates for finite-state Markov chains. Estimates have been characterized in terms of the log-Sobolev inequality (Diaconis and Saloff-Coste (1996a)), the Poincare inequality on graph representations (Diaconis and Strook (1991), Tuominen and Tweedie (1994)), and the spectral gap (Diaconis and Saloff-Coste (1996b), Saloff-Coste (2004)). Each of these build on Doob (1959), Nash (1958) and more recently Anderson (1989). Diaconis (2009) provides an excellent discussion of related developments. Fukushima (2010) emphasizes quadratic and bilinear form representations of Markov processes on general state spaces and touches on their spectral content. Chen, Hansen and Scheinkman (2007) make this connection explicit for the Feynman-Kac semigroup.

1.3 A Markov Model of Returns

We construct an empirical model to use for forecasting. Sections 3.1 and 3.2 describe the process environment and specify asset prices and a risk-neutral measure. Section 3.3 reviews the decomposition. Section 3.4 provides an example to highlight the economics of equilibrium prices and transitory fluctuations. 3.5 - 3.6 present the spectral gap, relates it to a martingale representation, and characterizes the identification of the spectral gap from the empirical covariance matrix.

The lag operator and composition of the lag and Markov operators are defined on the *path* space of the Markov chain, so some technical statements cannot be avoided. Extensive derivations are left for the section (7) appendix, along with the complete set of proofs.

1.3.1 Market Prices

Undiversifiable risk arises from S - state Markov jump dynamics taking values in a finite ordered set $x_t \in X := \{x^{[1]}, ..., x^{[S]}\} \subset \mathbb{R}^S$. Time is discrete. We take each $x^{[j]} \in \{0, 1\}$ so the state at time t, $x_t = x^{[j]}$ is characterized by the index $\{[j] : x^{[j]} = 1\}$. Local dynamics of the Markov chain X_t are described by the *kernel* function $m(x^{[i]}, x^{[j]}) := \Pr(X_t = x^{[j]} | X_{t-1} = x^{[i]})$.

Sequentially traded state-contingent securities $d_n \in F(X) \subset \mathbb{R}^S$ are sufficient to build up rich dynamically complete cross-sections as in Arrow (1953). We specialize asset $n = 0$ to $d_0 = (1, 1, ..., 1)'$.[5] We will extend the marketable security space to include long-lived securities recursively, but we first establish the benchmark asset prices.

Market equilibrium implies a positive pricing kernel exists and can be used in lieu of replication to price arbitrary cash flows (Ross (1976), Harrison and Kreps (1979)). Let $s_{j,t} = s_j(w_j(x_t), t) = \beta_j^t \, \tilde{s}_j(w_j(x_t))$ be the marginal value of wealth for investor j, where $\beta_j \in (0, 1)$ captures time discounting. For any asset n and market prices $p_{n,t}$, individual optimality requires

$$s_{j,t} p_{n,t} = \mathbb{E}_t[s_{j,t+1} d_n(X_{t+1})] \tag{3.1.1}$$

[5]Contingencies d_n are assumed to satisfy $\nu d_n d_n' < \infty$, which in finite dimensions under any positive probability measure ν is equivalent to $d_n \cdot e_i \leq M < \infty$ for every $n, 0 \leq i \leq S$ and some fixed scalar $M \in \mathbb{R}$, where $\{e_i\}, 0 \leq i \leq S$ denotes the standard basis in $\mathbb{R}^S$.

in each period t. In equilibrium, the stochastic discount factor (SDF) $S_{t,t+1} = S(X_{t+1}|x_t, 1)$ encodes market-wide preferences in observed prices, enforcing

$$s_{j,t}\mathbb{E}_t[S_{t,t+1}d_i(X_{t+1})] = \mathbb{E}_t[s_{j,t+1}d_i(X_{t+1})]$$

for any unconstrained investor j, and arbitrary i, t. In particular, 3.1.1 becomes

$$p_{n,t} = \mathbb{E}_t[S_{t,t+1}d_n(X_{t+1})] \tag{3.1.2a}$$

for every n given t and every t. We follow the convention in Alvarez and Jermann (2005) by modeling the SDF as the ratio of the pricing kernels $S_{t,t+k} = s_{t+k}/s_t = \beta^k \tilde{s}_{t+k}/\tilde{s}_t$. Then, the pricing kernel is the particular SDF when the reference period wealth is the numeraire $s_t = S_{0,t} = S(X_t|x_0, t)$, $s_0 \equiv 1$. For the asset $n = 0$, $p_{0,t} = \mathbb{E}_t[S_{t,t+1}] = 1/r_{f,t}$ is the price of a one-period default free bond per unit of face value.

1.3.2 Decomposition

A first-order transition distribution can be decomposed into two orthogonal components.[6] Each transition probability $m(j, k)$ is comprised of a local *transitory* and a non-local *permanent* component. The local components contains state-dependent transitioning information. The non-local component completely determines asymptotic forecasts. Both components are important in finite samples.

Let $\iota = 1_{S\times 1}$; then by our convention, $\mathcal{M}\iota = \iota$. We assume the chain $(X, \mathcal{M})$ is ergodic, which implies a unique invariant $\mu_0' = \mu_0' \mathcal{M}$. The pair (μ_0, ι) are the left and right eigenvectors of $\mathcal{M}$, respectively, normalized so that μ_0 is a probability measure $\mu_0'\iota = 1$. From

[6] See proposition (7.1) parts $I. - III.$ and corollaries (7.4) – (7.9)).

these assumptions, we obtain the representation for the dynamics of distributions over X

$$\mathcal{M}' = \mu_0 \iota' + \mathcal{M}'_{\boldsymbol{\gamma}}$$

Asset return dynamics inherit this representation

$$\begin{aligned}\mathbb{E}[R_{n,t+1}(X_{t+1})|x_{t,k}] &= 1(x_{t,k})'\mathcal{M}r_n \\ &= \overline{r}_n + 1(x_{t,k})'\mathcal{M}_{\boldsymbol{\gamma}} r_n\end{aligned}$$

where $\overline{r}_n := r_n \cdot \mu_0 = 1(x_{t,k})'(\mu_0 \iota')' r_n$ is the long-run mean return for asset n. The operator $\mathcal{M}'_{\boldsymbol{\gamma}}$ drives purely transitory variation.[7] In corollary (7.6), we establish the classic Wold representation applied to returns,

$$R_{n,t+1} = \overline{r}_n + r_n \cdot \sum_{s=0}^{\infty} (\mathcal{M}'_{\boldsymbol{\gamma}})^s 1(u_{t+1-s})$$

In lemma (7.2), we identify $\nu = \mathcal{M}'1(x_{\cdot,k})$ with probability measures over X given $x_{\cdot,k}$ for any k . Hence, the decomposition states that any transition probability can be written

$$p_{i,j} = \mu_j + \gamma_{i,j} \tag{3.3a}$$

for order pairs $x^{[i]}, x^{[j]}$. 3.3a indicates that conditioning on $x_{t,i}$ we arrive at $x_{t+1,j}$ with probability equal to the long-run occupation rate of the coordinate j plus a correction term $\gamma_{i,j}$. In corollary 7.1.2, we show that $\mathcal{M}_{\boldsymbol{\gamma}}\iota = 0$, i.e., $\sum_j \gamma_{i,j} = 0$. Hence, if any individual term $\gamma_{i,j}$ is nonzero for a given i, then at least one of the entries is negative for that i.

[7]We assume throughout that the columns of $\mathcal{M}'$ are not each identically μ_0, so the decomposition is nontrivial (i.e., $\mathcal{M}'_{\boldsymbol{\gamma}} \neq \mathbf{0}$).

1.3.3 The Spectral Gap

The *spectral gap* measures the difference in levels between the two largest eigenvalues of a Markov operator. The spectral gap provides a parsimonious statistic for finite sample deviations from stationarity. If asset returns are commensurate with Markov dynamics, the spectral gap measures the difference in the squared prices of risk associated with the positive-supply market factor and the next largest source of common variation.

Transition dynamics in general are distinct from dynamics of conditioning information, but can be characterized tractably. Using proposition (7.1) and corollaries (7.2)-(7.3), expected return dynamics are

$$\begin{aligned}\mathbb{E}_t[R_{t+k}] - \mathbb{E}_{t-1}[R_{t-1+k}] &= \Delta[(\mathcal{M}')^k x_{t-1}] \\ &= \Delta\left(\mathcal{M}'_{\boldsymbol{\gamma}}\right)^k x_{t-1} + \Delta\widehat{\mathbb{E}}_{t-1,k}\end{aligned}$$

The term $\Delta\left(\mathcal{M}'_{\boldsymbol{\gamma}}\right)^k$ captures changes in transition probabilities conditionally. Following Dyson (1962), or Tao (2010), the leading order terms for dynamics of $\left(\mathcal{M}'_{\boldsymbol{\gamma}}\right)^k$ are

$$\Delta[\left(\mathcal{M}'_{\boldsymbol{\gamma}}\right)^k] = -(1-\lambda_2)^{-k}\gamma\gamma' + o\left(\cdot\right) \tag{3.1s}$$

The kernel function γ of $\mathcal{M}_{\boldsymbol{\gamma}}$ can be viewed as the left eigenvectors of $I - \mathcal{M}_{\boldsymbol{\gamma}}$, even if $\mathcal{M}$ is not reversible. Lower eigenvalues of $\mathcal{M}$ are eigenvalues of $\mathcal{M}_{\boldsymbol{\gamma}}$. We show this in sections 7.4-7.5. Setting $\langle\gamma, 1\rangle = \boldsymbol{\gamma}$, and $\Delta\widehat{\mathbb{E}}_{t-1,k} = 0$ and conditioning,

$$\mathbb{E}_t[R_{t+1}] - \mathbb{E}_{t-1}[R_t] = -\zeta_t^{-1}\boldsymbol{\gamma}\gamma x_{t-1} \tag{3.2s}$$

describes time-varying mean returns. $\zeta_t = \lambda_1 - \lambda_2 = 1 - \lambda_2$ is the *spectral gap* of $\mathcal{M}$.

The dynamics of spectral data provide a summary for the dynamics of the risk prices because in equilibrium the risk prices are eigenvalues. We consider a fixed known covariance matrix plus mean zero i.i.d. random perturbations.

1.3.4 Identification with Empirical Covariance

The process for expected returns is described by the dynamics of the spectral gap. We identify the spectral gap from the PCA of the asset return variance-covariance matrix. We obtain two expressions. The first is from the Markov generator,

$$\mathbb{V}(R) = \mathsf{U} D_{1-\lambda} \mathsf{U}' \Sigma$$

where $D_{1-\lambda} := (I - \mathbf{\Lambda})^{-1}$ and $\mathbf{\Lambda}$ contains the singular values of data generated by $\mathcal{M}$. The second is from a singular-value decomposition (SVD) of observed asset returns,

$$\mathbb{V}(R) = \mathsf{V} \Lambda_{PCA} \mathsf{V}'$$

Equating expressions

$$\mathsf{V} \Lambda_{PCA} \mathsf{V}' = \mathsf{U} D_{1-\lambda} \mathsf{U}' C C'$$

where $CC' = \Sigma$ is the Cholesky decomposition of the covariance matrix of forecast errors $u_{t+1} = r(x_{t+1}) - (\mathcal{M}r)(x_t)$. For the benchmark $CC' = \Sigma = I$, pointwise identification can be stated

$$\zeta^{-1} = (D_{1-\lambda})_{2,2} = (\Lambda_{PCA})_{2,2}$$

We will use this ζ_t to define the changes of measure h. From here, we can represent the empirical time series of returns in the forms 3.2b, 3.2s and in the Wold form (corollary (7.6)).

We do not recover "true" physical transition dynamics from the identification as debated in Ross (2015) and Borovicka, Hansen and Scheinkman (2017), although clearly we use related machinery. Instead we recover incremental changes to the balance of deviations from the Perron-Frobenius limit. The technical identification of the Markov components from the

PCA of asset returns, developed in sections (7.3.1) and (7.4) of the appendix, provides the complete details.

1.4 Empirical Implementation

1.4.1 Nominal Forecasts

We construct *nominal* risk factor return forecasts within each rolling window for several horizons k separately. Nominal forecasts are fitted values of factor returns on past dividend yields and past returns. The sample window is truncated for each t depending on k, so that the effective windows for nominal forecasts are $\tau_m(t,k) := [t-(T_M-k),t]$. In the main test results reported, we reduce the window for all returns and forecasting series to that of the maximum horizon $\overline{k}$ nominal forecast series, $\tau_m(t,\overline{k})$. More nuanced procedures do not improve the performance significantly.

Nominal forecasts should not be confused with the forecasts made using the decomposition, which are the basis of the o.o.s. tests. The derived forecasts perform significantly better than the nominal forecasts used as inputs because the nominal forecasting procedures exploit none of the Markov structure of equilibrium asset returns.

1.4.2 Sample Window and Mixing

A distinction is made between the window size T_m and the mixing times $N(\epsilon_0)$ for threshold $\epsilon_0 > 0$. As a benchmark, the fixed window size T_m proxies for the stationary mixing times $N_t(\epsilon_0) = N(\epsilon_0, R_{t-T_m,t})$. The mixing time at t is defined

$$N_t(\epsilon_0) := \min_s \left\{ \mathbb{R}_+ \ni s \geq t : \max_{x \in X} \parallel \hat{h}_{t,t+s}(x) - \mu_0(x) \parallel \leq \epsilon_0 \right\}$$

In words, the mixing time is the shortest amount of time it can take for the maximum total variation distance between rows of the transition matrix to be within a threshold of size ϵ_0. Using the convergence results from proposition (7.1) part (II), we can bound the mixing time

$$N_t^*(\epsilon_0) = \log(\zeta_{0,t})^{-1}[\log(\epsilon_0) - \log(\chi_t)] \tag{4.2}$$

where $\chi_0\mu_0 = \hat{h}_{0,0}$ is the initial statistical likelihood ratio of the conditional to the unconditional distributions. Changes in the window size T_m are justified by changes in $N_t(\epsilon_0)$, which for fixed $\epsilon_0 > 0$ varies through estimates of χ_t and $\zeta_{0,t}$. We verify ex-post that we cannot distinguish the sequence of mixing times $N_t(\epsilon_0)$, $t \in [0, T]$ from a covariance stationary process.

It is also possible to choose the window size in each period $T_m = T_m(t)$ to minimize the ℓ^2- distance between the mixing time estimate for that period and a fixed target mixing time $N_0(\epsilon_0)$. Here, the mixing time estimator is formally a transformation of a random sample. This objective corresponds to a well-defined extremum estimator. However, empirically we find this step produces very little movement in the window size $T_m(t)$ over t.

1.4.3 Data

For priced risk factors, we use monthly returns data on the Fama-French three-factor, Fama French three factor plus momentum (Carhart), and Fama-French 5- factor models. The Fama-French three factors are the Market, Value (HML) and size (SMB), rebalanced annually. Details are in Fama and French (1993). Momentum is constructed from a portfolio long 2-12 month winners and short 2-12 month losers ranked in the cross-section and truncated at the 30% and 70% percentiles. Each reported momentum sorted portfolio is an average of small and large-cap momentum stocks. Momentum is rebalanced monthly.

We consider several cross sections of test assets including the 25 size-BTM portfolios,

the 25 size-BTM plus 10 Momentum portfolios, the 25 size -operating profit portfolios and the 32 size-BTM-OP portfolios. The size-BTM portfolios are annually rebalanced and are comprised of the intersection between five market-cap sorted stocks with five book-equity to market-equity (BTM) sorted stocks. The size-BTM plus Momentum add 10 portfolios sorted on 2-12 month returns cross-sectional rankings. The size-OP portfolios sort annually on operating profits (OP): "annual revenues minus COGS, interest expense, and SG&A", normalized by trailing book-equity, and intersect with size. The factor returns data and the test asset returns data are obtained from Ken French's website.[8]

We use predictor variables from Goyal and Welch (2008). Data are available on Amit Goyal's website.[9] We use monthly data for the rolling average 12-month dividends, the rolling average 12-month earnings, and the index level for the $S\&P500$. Data are from 1926-2016. Monthly value-weighted and equal- weighted index total returns and ex-dividend returns over 1926-2016 are from CRSP. 18 portfolios sorted by cash flow to market capitalization and 18 sorted by dividend to market capitalization over the same period are from Ken French. We form three high-low portfolios for cash flow and three for dividends, leveraging the cash flow spread at denary, quinary and tertiary scales. The breakpoints are available on Ken French's website.

1.5 Empirical Results

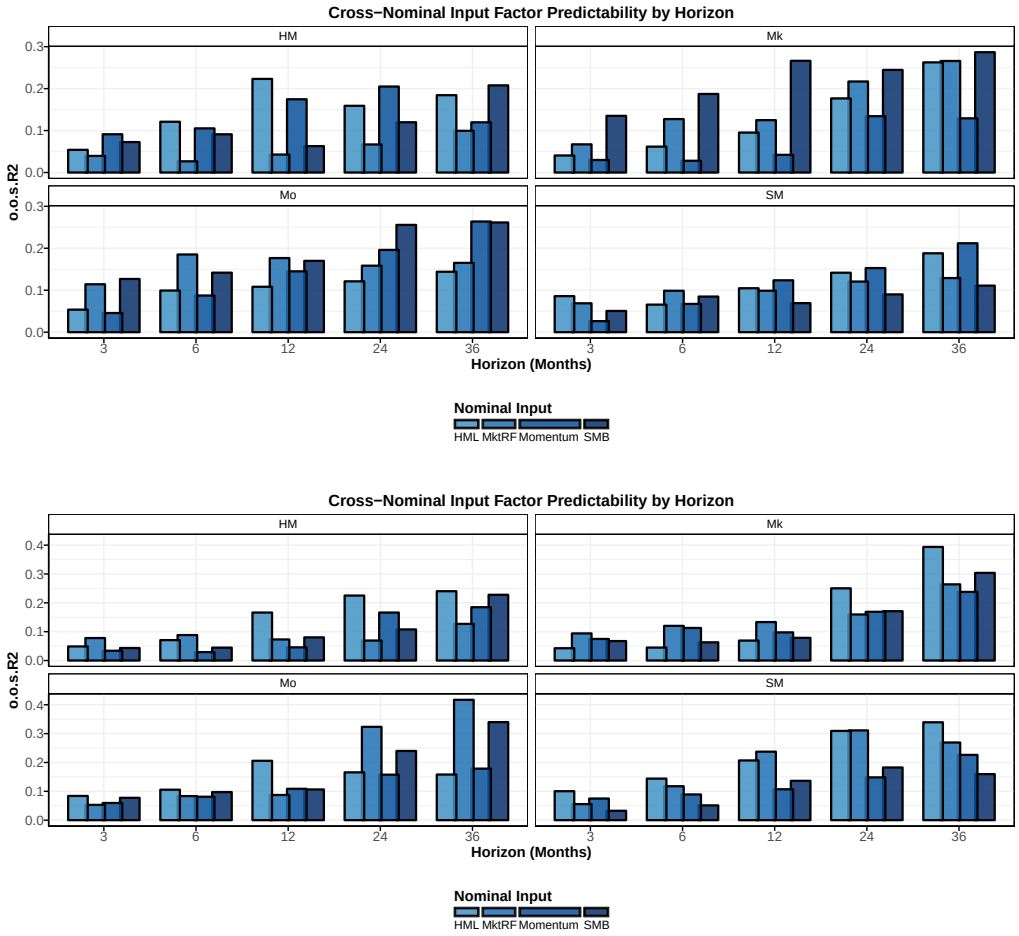
1.5.1 Latent Factor Dynamics

Persistence estimates for the leading and penultimate latent factors are reported in Table 1. We report coefficient estimates from a first-order autoregression for the demeaned factor processes. The leading factor exhibits no predictable deviation from its mean. In contrast,

[8] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

[9] http://www.hec.unil.ch/agoyal/

Figure 1.2: Predictability Term Structures for the Carhart Factors



(a) Out of sample forecast performance for each of the Carhart four factors. Value is forecasted by the volatility of the spectral gap. The spectral gap is calculated up to time t using data from $[T_m - t, t]$ and used to forecast returns for various $t + k$, including returns between periods $t + 1$ and $t + n + 1$ for $n = 1, 2, 4, 8, 12$. Top panel has $T_m = 8$yrs. Lower panel has $T_m = 12$yrs. Carhart Factors are the Fama-French Market, Value and Size factors plus the Momentum factor (UMD). Quarterly data $Q1$ 1967 to $Q3$ 2015 are compounded monthly factor returns from Ken French's website.

the penultimate factor is highly persistent. Fluctuations around the mean of the penultimate factor are predictable, stationary and statistically significant. GMM standard errors adjust for serial correlation. The sharp contrast in predictability between the leading and penultimate factors is consistent with the predictions of the Markov asset pricing model with nonzero stationary mixing times.

Dynamics of the factors of realized returns are plotted in Figure 2. Figure 2 in combination with Table 1 restate the findings in Alvarez and Jermann (2005) that the bulk of pricing kernel variation comes from the permanent factor. Transitory factors are plotted in Figure 3, which shows the time series of the time-varying expected return factors. Figure 1 charts the empirical densities of conditional risk prices. Figure 4 plots the dynamics of the empirical spectral gap. The spectral gap is strongly countercylical. The interpretation of Figure 4 is that *volatility increases in bad times, but so does the concentration of volatility on the permanent factor.*

Latent and Conventional Factor Composition

The martingale factor captures 84% of the common time series variation in asset returns. The market and the martingale factor are almost identical. The conventional value (HML) factor provides a striking contrast to the market factor. Table 8 breaks down the market and HML loadings on the latent expected return components. Variation in returns to the value factor load significantly on the leading expected returns factor, which drives transitory predictable fluctuations in mean returns. The conventional market factor does not significantly load on the expected return factors.

Table 1.1: Market Return Predictability by Dividend Yield, *cay* and the Spectral Gap

Panel I shows out of sample predictability of market returns by the spectral gap. The spectral gap is the difference between the conditional volatilities of the permanent and first transitory factors, measured by the second conditional eigenvalue of the empirical decomposition of asset returns. $II.$ shows the out of sample predictability for the dividend yield. Panel $III.$ shows out of sample prediction statistics for *cay*. The spectral gap *excluding* non-market volatility is given in panel IV. The 12-month moving average of monthly dividends, the market index level and *cay* are from Goyal and Welch. Fama and French factor returns quarterly are from Ken French. Data are quarterly from 1967 $Q1$ to 2015 $Q3$.

		Full Sample Estimates		Out of Sample Statistics	
Predictor	k (quarters)	coefficient $\widehat{\phi}_{j,k}$	$t : H_0(0)$	R^2	adj.R^2
$I.$	1	2.39**	2.770	0.038	0.033
Spectral	2	4.51**	3.650	0.065	0.060
Gap	4	8.01***	4.805	0.108	0.104
	8	14.76***	6.893	0.203	0.199
$II.$	1	64.066	1.256	0.008	
Dividend	2	143.25*	1.929	0.019	
Yield	4	243.69*	2.385	0.029	
	8	355.27**	2.555	0.034	
$III.$	1	47.51*	1.944	0.019	
cay	2	101.01**	2.821	0.040	
	4	196.51***	3.974	0.076	
	8	388.12***	5.901	0.157	
$IV.$	1	1.48*	2.136	0.023	
Market Gap	2	2.90**	2.895	0.042	
	4	5.29**	3.896	0.074	
	8	9.15***	5.095	0.122	

Table 1.2: Persistence of first and second components of realized returns.

The first component is not predictable, while the second and trailing components are predictable when expected returns are time-varying. For the j'th component of returns we report the estimated autoregressions

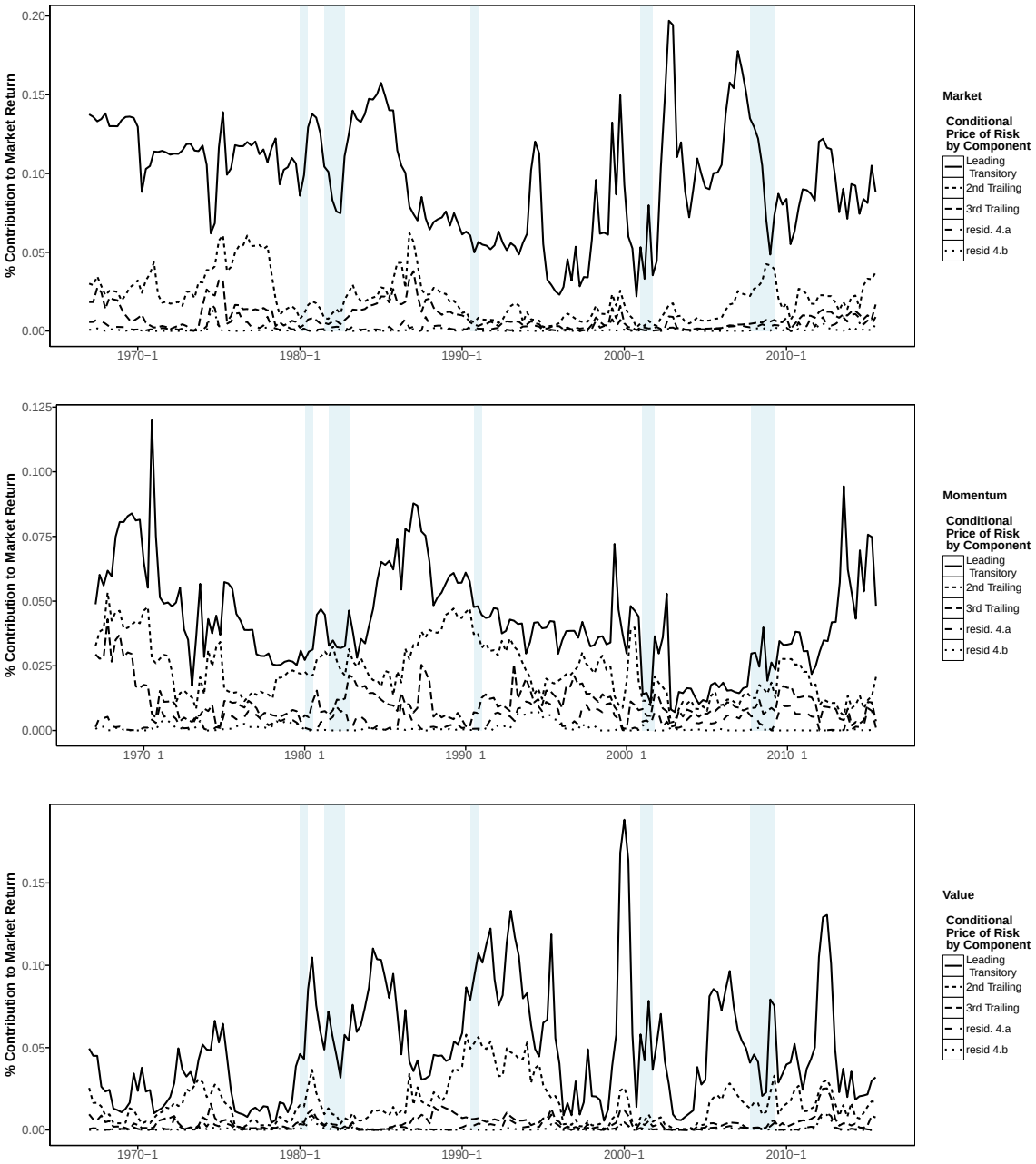
$$x_{t,j} = \phi_{0,j} + \phi_{1,j} x_{t-1,j} + u_{t,j}$$

The components are obtained over rolling $T_m = 15$ year samples of realized returns supplemented with the forecast errors from dividend yields over several horizons. A singular value decomposition of the generalized covariance matrix orders orthogonal components by contribution to variation in realized returns and lower frequency variation from surprises to historical forecasts. We report autoregressions over the full sample for components $j = 1, 2$.

	coefficient	estimate	$s.e$
Leading Factor	$\widehat{\phi}_{0,1}$	0.312	0.624
	$\widehat{\phi}_{1,1}$	0.089	0.072
Trailing Factor	$\widehat{\phi}_{0,2}$	-0.059	0.032
	$\widehat{\phi}_{1,2}$	0.649^{***}	0.055

(a) The penultimate component of the realized returns is the leading component of expected returns. Returns data are quarterly from $Q1$ 1967 to $Q3$ 2015. Fama French and Carhart factor model returns are from Ken French's website.

Figure 1.3: Time Varying Components for the Market, Momentum and Value



(a) Dynamics for the components of common variation for the market, momentum and value expected returns, expressed as a percentage of the total variation of factor returns. Expected return variation comprises roughly 20% of the variation in total returns. Expected return variation comprising market returns contribute an average of less than 14% of variation in total returns, and comprising momentum returns contribute an average of less that 5% of the variation in total returns. Quarterly data $Q1$ 1967 to $Q3$ 2015. Fama French and Carhart factor model returns are from Ken French's website. NBER recessions are in blue.

Table 1.3: HML Return Predictability

(a) Value return predictability, assessed over the entire sample using a time-series of forecasts made out of sample. The predictive regression is

$$R_{t+k,j} = a_0 + a_1 G_j(\widehat{\zeta_t}) + \varepsilon_t$$

where $G_j(\hat{\zeta}_t \in \{\sigma_1/\overline{\sigma}, \sigma_2\}$. The rolling T_m samples are augmented with information contained in the nominal forecast errors at lags 2^k, $k \in [0,5] \cap \mathbb{N}$. Nominal inputs are listed. Data are from $Q1$ 1967 to $Q3$ 2015 from Ken French, Goyal and Welch (2008) and CRSP.

		Full Sample Estimates			Out of Sample Statistics	
Nominal Input	horizon (quarters)	predictor	estimate	$t : H_0(0)$	R^2	adj.R^2
DP &	1	σ_2	4.97	1.339	0.009	0.004
10-1 CF		$\sigma_1/\overline{\sigma}$	-7.37	-1.272	0.008	
Portfolio	2	σ_2	24.29 ***	4.505	0.096	0.092
		$\sigma_1/\overline{\sigma}$	-38.41 ***	-4.573	0.099	
	4	σ_2	40.14 ***	5.229	0.127	0.122
		$\sigma_1/\overline{\sigma}$	-67.46 ***	-5.702	0.147	
	8	σ_2	46.41 ***	4.31	0.092	0.087
		$\sigma_1/\overline{\sigma}$	-80.96 ***	-4.887	0.115	
	12	σ_2	46.15 **	3.830	0.075	0.070
		$\sigma_1/\overline{\sigma}$	-84.34 ***	-4.560	0.104	
DP	1	σ_2	28.84	1.501	0.012	0.006
		$\sigma_1/\overline{\sigma}$	-79.50	-1.427	0.010	–
	2	σ_2	125.66 ***	4.508	0.097	0.092
		$\sigma_1/\overline{\sigma}$	-370.56 ***	-4.598	0.100	–
	4	σ_2	191.03 ***	4.768	0.108	0.103
		$\sigma_1/\overline{\sigma}$	-576.42 ***	-4.993	0.117	–
	8	σ_2	214.77 **	3.831	0.074	0.069
		$\sigma_1/\overline{\sigma}$	-666.78 ***	-4.131	0.085	–
	12	σ_2	122.28 *	1.941	0.020	0.015
		$\sigma_1/\overline{\sigma}$	-383.66 **	-2.107	0.024	–

1.5.2 Forecasting

We use the common sources of transitory variation to predict factor returns out of sample. Results are reported in Table 2. The spectral gap (lagged for predictive regressions) forecasts the market returns with an o.o.s. R^2 of 3.8% quarterly, 6.5% semiannually, 10.8% annually and 20.3% biannually. Other variants of the spectral gap, including the common negative exponential transform, perform similarly but no better. *cay* performs relatively well, giving an out of sample R^2 of 7.6% on an annual basis, while the dividend yield generates forecasts with o.o.s. R^2 of only 2.9%.

This latter number contrasts with conventional wisdom because it is an out-of-sample estimate. The original predictability studies by Campbell and Shiller (1989) were estimated in sample (see Goyal and Welch, 2008). The roughly 10% R^2 reported by Cochrane (2008) is calculated using the Campbell-Shiller decomposition, which jointly restricts cash flow and discount rates by design. The forecasts in Cochrane (2008) are necessarily evaluated in sample using fitted values from a vector-autoregression (VAR).

In fact, our evidence corroborates the importance of an aggregate measure of dividend yields for forecasting risk premiums. While the short-horizon performance of the dividend yield is poor out of sample, explained variation grows monotonically, resulting in the stylized fact that low yields today (high prices) are followed by low returns tomorrow. As emphasized by Cochrane (2011), variation in the dividend yield corresponds entirely to variation in mean returns (at the aggregated level), and more so at the long end.

Our benchmark nominal forecast variable is the dividend yield for the CRSP value-weighted index. We construct a rich information space for our decomposition by calculating optimal return forecasts using the dividend yield, separately for lags of 2^k, $k = 0, 1, 2, ..., k^+$. Surprises from different lag lengths are calculated separately. An $AR(1)$ generates signals that are all deterministic functions of each other, eliminating signals relevant over different time scales when they exist.

We also find that constructing nominal forecasts using the CRSP ex-dividend value-

weighted index return translates into significantly poorer performance. Value in this case cannot be forecasted. The explained variability in market returns does not exceed 10% out of sample at any horizon in this case. Interestingly, an equally-weighted dividend yield used to generate nominal forecasts also results in diminished efficacy. Unsurprisingly, size is hit least by this distinction. At the three-year horizon, the value weighted dividend yield nominal input predicts 34.7% of the variation in returns to the SMB portfolio. Using the equal -weighted yield generates an R^2 of 30.5%.

Several features of the value forecasts reported in Table 3 are striking. First, at the annual horizon, the variation in value (HML) returns is 14.7% predictable, and semiannually, 9.9% predictable. Because HML is a priced risk factor, these percentages measure the fraction of the variability in value premia that are predictable. Said another way, they capture the time-varying price of risk for exposure to innovations in the HML replicating portfolio. Second, unlike market predictors, which, as emphasized by Cochrane (2008, 2011), increase in explanatory power monotonically with horizon, the value predictors have a half-life of about a year. In every case, the two year prediction is less effective than the one year prediction.

The shape of the term structure of the time-varying value premium is robust to the choice of predictor as long as the predictor works. We report two normalizations of the second moment of the spectral gap that forecast HML returns. Every other priced risk factor we forecast has an upward sloping term structure of predictability, making the value premium unique among priced sources of risk in U.S. equity markets. This finding is also consistent with the findings in Koijen, Lustig and Van Nieuwerburgh, showing value returns load on the transitory variation picked up by the Cochrane-Piazzesi factor.

From Table 4. the momentum factor returns are predictable. The first two years explained variation is low, but accelerates after the second year to meet or exceed market predictability after the 3-year horizon. Momentum predictability jumps from 6.4% at one-year to 16.1% and 31.4% at the two- and three- year horizons, respectively. The curvature

Table 1.4: Market Return Predictability

(a) The spectral gap predicts market returns. Predictability regressions are calculated over the entire sample. The predictor series $\hat{\zeta}_t$ is constructed using data up to t for forecasts of returns at $t+k$ or $t+1+k$, $k \in \{1, 2, 4, 8\}$ on rolling windows of T_m =15yr histories. The predictive regression is

$$R_{t+k,j} = a_0 + a_1 G_j(\widehat{\zeta}_t) + \varepsilon_t$$

where $G_j(\hat{\zeta}_t)$ is a functional of the spectral gap time series that can vary only by portfolio, indexed by j. The rolling T_m samples are augmented with information contained in the forecast errors from forecasts made at various lags. The variable $\widehat{\zeta}$ is shorthand for the empirical estimate of the spectral gap. Tr_1 is the (non-normalized) largest contribution to the trace of the covariance matrix. Monthly dividend yields, Fama -French 3-factor and Carhart model returns and *cay* data are from $Q1$ 1967 to $Q3$ 2015 from Ken French, Goyal and Welch (2008) and CRSP.

		Full Sample Estimates			Out of Sample Statistics	
horizon (quarters)	target	predictor	estimate	$t : H_0(0)$	R^2	adj.R^2
1	R_M	$\widehat{\zeta}$	2.546	2.983	0.044	
		$(Tr)_1$	1.354	1.893	0.018	–
		$\widehat{\zeta} + (Tr)_1$	1.169	2.683	0.036	0.031
2	$R_{M,2}$	$\widehat{\zeta}$	3.827	3.065	0.047	
		$(Tr)_1$	2.357	2.257	0.026	
		$\widehat{\zeta} + (Tr)_1$	1.879	2.956	0.044	0.039
4	$R_{M,4}$	$\widehat{\zeta}$	7.708	4.486	0.098	
		$(Tr)_1$	4.682	3.249	0.054	
		$\widehat{\zeta} + (Tr)_1$	3.778	4.320	0.091	0.086
8	$R_{M,8}$	$\widehat{\zeta}$	13.181	5.626	0.151	
		$(Tr)_1$	7.013	3.568	0.067	
		$\widehat{\zeta} + (Tr)_1$	6.167	5.142	0.129	0.124
12	$R_{M,12}$	$\widehat{\zeta}$	18.307	6.480	0.198	
		$(Tr)_1$	9.032	3.791	0.078	
		$\widehat{\zeta} + (Tr)_1$	8.494	5.800	0.165	0.160

of momentum predictability is increasing and convex and in this way stands out from the nearly linear term structure of market predictability. Standard market predictors do poorly with momentum, reported in Table (10). Earnings to price, cay and the dividend yields all predict less than 3% of the variation in momentum return at any horizon.

Figure (1) contrasts the term structures of return predictability for the value (HML), Momentum (UMD), market and size (SMB) factor replicating portfolio returns. Explained variation grows monotonically in forecast horizon for the market, momentum and size. Time-varying expected returns to value are predictable in the short-run, but become negligible in asymptotic forecasts. The market term structure is nearly linear. The momentum term structure is a locally increasing convex function of horizon while value predictability is a concave function with a local maximum at the one-year horizon. Forecasts are given by the lagged spectral gap normalized to match the fit of the realized factor returns on rolling historical $T_m = 15$ year samples.

These forecasts meet or exceed existing predictors in the literature to our knowledge, with the exception of Kelly and Pruiit (2013) in the case of market returns. Kelly et. al find an o.o.s R^2 of 13% for the market, omitting windows in a neighborhood the forecast target.[10] In the case of momentum returns. Huang (2016) reports a monthly out of sample R^2 of 0.5% for momentum returns using the cross-sectional dispersion of moving average annual returns (the "momentum gap"). This compares roughly to the spectral gap's forecast R^2 of 6.4% when scaled to the one-year horizon.

Using the CRSP value-weighted index ex-dividend as the nominal input produces limited forecasting power but has interesting implications for the market. The market o.o.s. R^2's of 2.8% and 6.9% at the 1 - and 2 -year horizons accelerate to 16.3% at 3-years. At 3 years we report a significant time-series coefficient point estimate 15.54 with t- statistic 5.76%. This suggests capital gains have a non-negligible role forecasting aggregate returns over longer

[10]Using future observations implies these forecasts cannot be implemented in real time. However, they may be a better measure of the forecasting ability of a theoretical investor within the model. Kelly et. al can also be implemented in non-Markovian settings.

Table 1.5: Size (SMB) and Momentum (UMD) Out of Sample Predictability

(a) Predictability regressions are calculated over the entire sample. The predictor series $\hat{\zeta}_t$ is constructed using data up to t for forecasts of returns at $t+k$ or $t+1+k$, $k \in \{1, 2, 4, 8, 12\}$. The predictive regression is $R_{t+k,j} = a_0 + a_1\widehat{\zeta}_t + \varepsilon_t$, where $\widehat{\zeta}_t$ is the spectral gap time series. The rolling $T_m = 15$yr samples are augmented with nominal forecast errors from forecasts made at various lags. $Q1$ 1967 to $Q3$ 2015 data from Ken French, Goyal and Welch (2008) and CRSP.

			Full Sample Estimates		Out of Sample Statistics
Portfolio	quarters k	variable	$\hat{a}_1$	$t : H_0 = 0$	R^2
Size	1	$\widehat{\zeta}$	1.067	2.252	0.026
(SMB)	2	$\widehat{\zeta}$	1.819	2.787	0.039
	4	$\widehat{\zeta}$	4.175	4.507	0.098
	8	$\widehat{\zeta}$	8.495	6.363	0.185
	12	$\widehat{\zeta}$	12.874	7.485	0.248
Momentum	1	$\widehat{\zeta}$	-0.544	-1.305	0.009
(UMD)	2	$\widehat{\zeta}$	-1.207	-2.070	0.022
	4	$\widehat{\zeta}$	-2.939	-3.566	0.064
	8	$\widehat{\zeta}$	-6.209	-5.854	0.161
	12	$\widehat{\zeta}$	-10.395	-8.820	0.314

horizons. Using the equally weighted CRSP ex-dividend index to construct nominal forecast inputs unsurprisingly allows our model to forecast size, although not quite as well as with nominal dividend yield inputs. The former and latter 3-year o.o.s. R^2 for size are 30.05% and 34.7% respectively.

We reproduce the analysis shifted forward an extra period to address potential concerns about systematic measurement error. If prices are measured with error in a persistent direction then returns from contiguous intervals are spuriously correlated. This problem is not as likely in liquid stock markets as over the counter or emerging markets contexts. This is confirmed in table 10 for the case of momentum returns, where the annualized point estimate for forecasts from $t+1 \mapsto t+1+k$ is equal to the point estimate obtained from $t \mapsto t+k$ up to one significant digit. We calculate the staggered forecasts for each of the portfolio returns we analyze and find no significant discrepancies.

The patterns of predictability extend to the full sample of available U.S. equity data, beginning in 1926 when the NYSE emerged as a dominant national exchange, among other reasons.[11]. The term structure of momentum is convex and increasing beyond the 1-2 year horizon. Value is concave and decreasing after the 1-year horizon, both as in the primary sample. The long-end of the value term structure decreases more slowly and from a slightly lower rate of predictability, maximized at an out of sample R^2 of 11.2%.

Cochrane (2011) emphasizes the importance of measuring the volatility of the expeted return estimates in addition to the standard errors of the estimators. Below, we report a test case for the market out of sample expected expected return series by calculating the volatility over the sample. As in Cochrane, the volatility is the same order of magnitude as the level. The tradeoff is smoother here than Cochrane's (2011) in sample VAR estimates, generating a Sharp ratio of over 0.7 in each test case.

[11]Brown, Mulherin, and Weidenmier (2006) discuss the pre-1926 history of the stock exchange industry in the U.S.

Table 1.6: Means and volatilities of Market forecasts under 4 regimes: Full sample 1967-2017 and pre-crisis 1967-2006 periods, under Carhart and CRSP HLCF earnings spread nominal inputs.

Model	Variable	Standard Deviation	Mean	Forecast Sharpe Ratio
Carhart Full	Forecast	9.924	7.084	0.714
	resid.	13.786	–	
Carhart Pre-Crisis	Forecast	9.742	7.084	0.727
	resid.	13.915	–	
HLCF Full	Forecast	9.410	7.084	0.753
	resid.	14.142	–	
HLCF Pre	Forecast	9.064	7.084	0.782
	resid.	14.366	–	

1.5.3 Carhart Model Factor Replication

The timing portfolio returns for a particular factor are not identical to the sequence of conditional forecasts obtained for that factor. This is because the conditional forecasts are constructed by exploiting information in the errors from past nominal forecasts that are not replicable in the space of the factors' test assets. Nonetheless, a projection of the conditional forecasts onto the test asset space improves the efficiency properties of the factor replicating portfolios.

The timing portfolio for HML returns can be split into two components. The raw HML Sharpe ratio over the full sample is 0.177, while the first HML timing component achieves 0.230 and the second component achieves 0.283. Results are stated in Table 5. The hybrid timing component replicates HML returns but with lower volatility, producing a Sharpe ratio of 0.301 over the full sample. HML does better prior to the 2007 financial crisis with a Sharpe ratio of 0.190. The timing portfolio components pre-crisis generate Sharpe ratios of 0.259 and 0.223 for the penultimate and trailing components, respectively. The overall timing portfolio does less well excluding the crisis, but we still detect a significant improvement. The trailing component - both the trailing factor and the weight of the HML replication on that factor - drives performance of the HML timing portfolio during the financial crises.

When several components are predictable, a simple diversification argument suggests the hybrid returns will be more efficient than either of the individual component replicating returns, which is part of what we see here. This intuition is delicate: theory tells us that investors are willing to sacrifice some mean-variance efficiency for inter-temporal hedging opportunities represented by the HML factor. We should expect to see a diversification benefit from combining the components within the HML factor, as we do, but it is not obvious what to expect combining components from different factors with the market.

The momentum factor timing portfolio is concentrated on a single expected return factor. The Sharpe ratio rises from 28% to 32.8%. The Sharpe ratios, means and volatility are significantly estimated. Moreover, we reject the null hypothesis that the difference in the Sharpe ratios is zero. Results are reported in Table 5.

The SMB timing portfolio represents an improvement in mean-variance efficiency of 63%. Statistics are reported in Table 5, along with the standard factor replicating portfolio statistics for reference. Interestingly, the incremental gains in efficiency do not arise from more precise estimates of the conditional price of risk for the size factor. This and related findings are the subject of the proceeding section.

1.5.4 Cross-sectional Implications

We find some of the returns to the timing portfolios provide additional cross-sectional pricing power. This is consistent with the view that the efficiency gains arise through a better accounting of time variation in risk prices. Full sample tests are implemented using Fama-MacBeth, with the expected return factor-based timing portfolio returns treated as risk factors.

In Table 6, we report cross-sectional pricing implications for the market and value timing portfolios. Prices of exposure to returns on the market timing portfolio are high and significant. α's are indistinguishable from zero. Residual variation in the market is not

Table 1.7: Sharpe Ratio Comparisons
Value, Momentum, and Size Timing Portfolios

Top: Full sample and pre- 2007 financial crisis Sharpe ratios and t- statistics for HML and HML timing portfolio by component. Statistics correspond to pre-crisis estimates. Mid: Sharpe ratios for momentum, the first two components of the expected return factors weighted corresponding to their contribution to momentum, and the momentum timing strategy. Lower panel: size, size timing portfolio and market Sharpe ratios. Test assets are the Fama-French FF25 Size/BTM plus 10 momentum portfolios. Factor data are the FF3 plus Momentum factor returns. Data are quarterly from 1927 $Q1$ to 2015 $Q3$.

Strategy	Value	Penultimate	Trailing	Value Timing
Monthly SR				
Full Sample	0.177	0.230	0.283	0.301
Pre-2007	0.190	0.259	0.223	0.243
		(3.403)	(2.920)	(3.191)
Strategy	Momentum			Momen. Timing
Monthly SR				
Full Sample	0.2801	0.3283	0.3009	0.3284
				(4.055)
Strategy	Size			Size Timing
Monthly SR				
Full Sample	0.1352			0.2138
				(2.941)

Table 1.8: Cross-sectional Pricing for Market and Value Timing Portfolios

In the Fama-French three and Carhart factor model test asset spaces, the Market timing portfolio contains the pricing power of the market portfolio. The residual variation MKTres is not priced. Value remains priced in both economies. In the Fama-French three factor model cross-section, the value timing portfolio is priced and so is the residual variation in the value factor. Market in this case is not priced. The timing portfolios are projections of the out of sample forecasts of factor returns made using the spectral data pertaining to a Markov operator. The Markov spectra are identified from an empirical PCA of asset returns augmented with linear combinations of past forecast errors that render the effective transition dynamics "memoryless."

coefficient	$\widehat{\lambda}_M$	$s.e.$	$t-$stat	$\widehat{\lambda}_V$	$s.e.$	$t-$stat	Test Economy
α	0.446	1.592	0.280				FF3 + Market Timing
Market timing	9.711	2.514	3.862				
HML	1.307	0.426	3.066				
Mktres	3.618	1.544	2.344				
SMB	0.814	0.506	1.609				
α	0.978	1.738	0.563				Carhart + Market Timing
Market timing	7.989	2.627	3.042				
HML	1.394	0.405	3.444				
Mktres	2.721	1.790	1.520				
UMD	0.662	1.862	0.355				
SMB	0.771	0.504	1.529				
α				2.293	1.702	1.348	FF3+ Value Timing
Value timing				9.982	2.507	3.982	
HMLres				-0.486	0.182	-2.671	
MktRF				-0.007	1.593	-0.004	
SMB				0.860	0.508	1.692	

Coefficients are prices of risk. Factors include the market residual Mktres, HML and SMB (top panel) and the market residual, HML, SMB, and Momentum (lower panel). UMD is "up minus down" for momentum portfolios long on previous year's winners and short previous year's losers, both in cross-sectional ranking. Test assets are the Fama-French FF25 Size/BTM portfolios, with momentum portfolios in the lower panel. Robust standard errors are Newey-West via GMM. Both standard errors and $t-$ stats are reported for convenience. Factor data are the FF3 factor returns. Data are quarterly from 1927 $Q1$ to 2015 $Q3$.

significantly priced when separated from the market timing portfolio, and its effect vanishes entirely with inclusion of the momentum factor. A similar effect is true for the value timing portfolio, although the residual variation is not driven out by inclusion of momentum (not reported). Residual variation in returns is defined as variation in the residuals of an OLS regression of the factor returns on its corresponding timing portfolio.

1.5.5 Related Econometric Methods

This section provides more detail in the context of contrasting similar decompositions relying on Principal component analysis (PCA).

PCA

Principal component analysis (PCA) can be implemented on a rolling basis to produce out-of-sample predictions. A classical linear PCA of the return variance-covariance matrix produces the representations

$$\mathbb{V}_t(R) = \widehat{V}_t D_t \widehat{V}_t'$$

where the t indicates we consider PCA calculated on rolling windows for each t. Two important considerations distinguish our work from these representations. First, we identify maps from $\mathbb{V}_t(R)$ to the second moments of a Markov operator, which we regard as generating dynamics of asset returns in equilibrium. We denote the outcome of this map $\mathbb{V}_t(R) = V_t \Lambda V_t'$. Second, to hold mixing times a fixed distance from zero, we build a model of fixed length

T_m, written $\mathscr{V}\mathscr{D}\mathscr{V}'$, constructed as follows

$$
\begin{aligned}
&\mathscr{V}_{t,t}\mathscr{D}_{t,t}\mathscr{V}'_{t,t} = V_{t,t}\Lambda_{t,t}V_{t,t} \\
&\mathscr{V}_{t,t-1}\mathscr{D}_{t,t-1}\mathscr{V}'_{t,t-1} = V_{t-1,t-1}\Lambda_{t-1,t-1}V'_{t-1,t-1} \\
&\vdots \\
&\mathscr{V}_{t,t-(T_m-1)}\mathscr{D}_{t,t-(T_m-1)}\mathscr{V}'_{t,t-(T_m-1)} = \\
&\quad V_{t-(T_m-1),t-(T_m-1)}\Lambda_{t-(T_m-1),t-(T_m-1)}V'_{t-(T_m-1),t-(T_m-1)}
\end{aligned}
$$

At time t, each of the T_m entries, call them $j \in [t-(T_m-1), t]$, come from the innovation term in the rolling window PCA at time j, indexed by (j, j) to indicate the front-most entry in the component series obtained at time j.

Projection onto the empirical bases implies the exact multi-variable representation for each scalar realization of the time-t returns

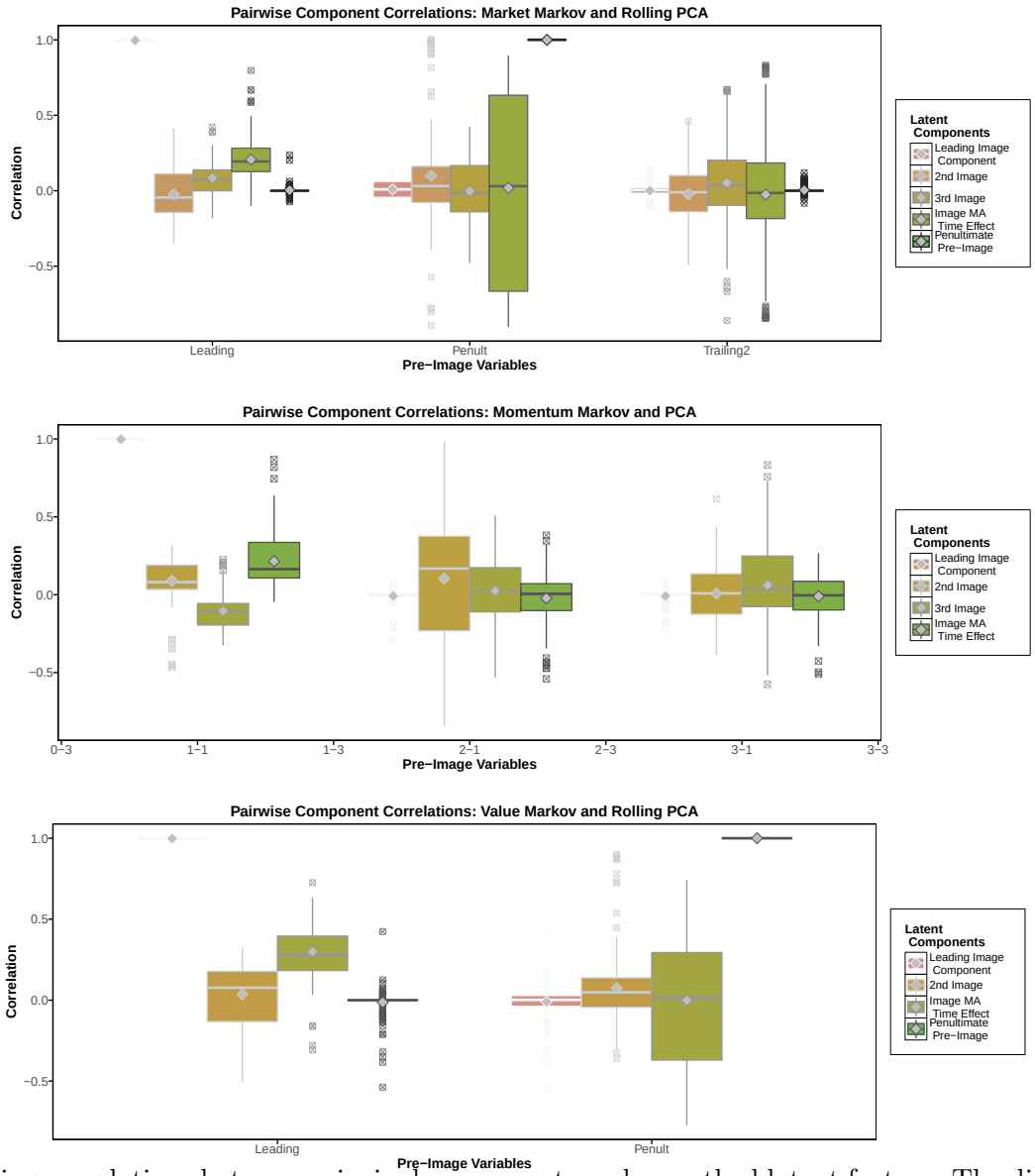
$$
R_{k,x,t} \;= \widehat{q_{0,k}}(x,t) + \sum_{n=1}^{N_0} \widehat{q_{n,k}}(x,t) \tag{3.3a}
$$

where $\widehat{q}$ are the orthogonal components, with in particular $\widehat{q_n}$ for $1 \leq n \leq N_0$ corresponding to weighted columns of V_t. $N_0 \leq N$ is the dimension of the representation.

Each period we produce a decomposition of variance over $T_M = 15$- year histories but use the contemporaneous entry for our representation. The distinctions between a companion rolling PCA arise from changes in the time-effect term in each period, and from cross-interacting between the components over medium-term histories. These interactions are prohibited by a classical PCA analysis.

A comparison of a rolling, smoothed PCA our model is illustrated in figure 7. The sequence of contemporaneous decompositions is more responsive to compositional changes associated with rare and asymmetric events. This can be seen in the form of violations of orthogonality between components across the two models. Components of the fixed mixing

Figure 1.4: Outlying Realized Correlations



(a) Rolling correlations between principal components and smoothed latent factors. The differences between the two are highlighted by the outlying correlations between the PCs and the Markov bases. Downturns in the business cycle are marked by jumps from zero to near one, in absolute value, in the correlation between the trailing components of the two models.

model become extremely correlated with cross-components in the PCA model during outlying and asymmetric events, while the components within each model are forced to stay orthogonal, missing some of the news content in the data. As suggested, the sources of this responsiveness are the time-effect term, and the flexibility of allowing components historically to be correlated but contemporaneously orthogonal.

1.5.6 Gains in Dimension

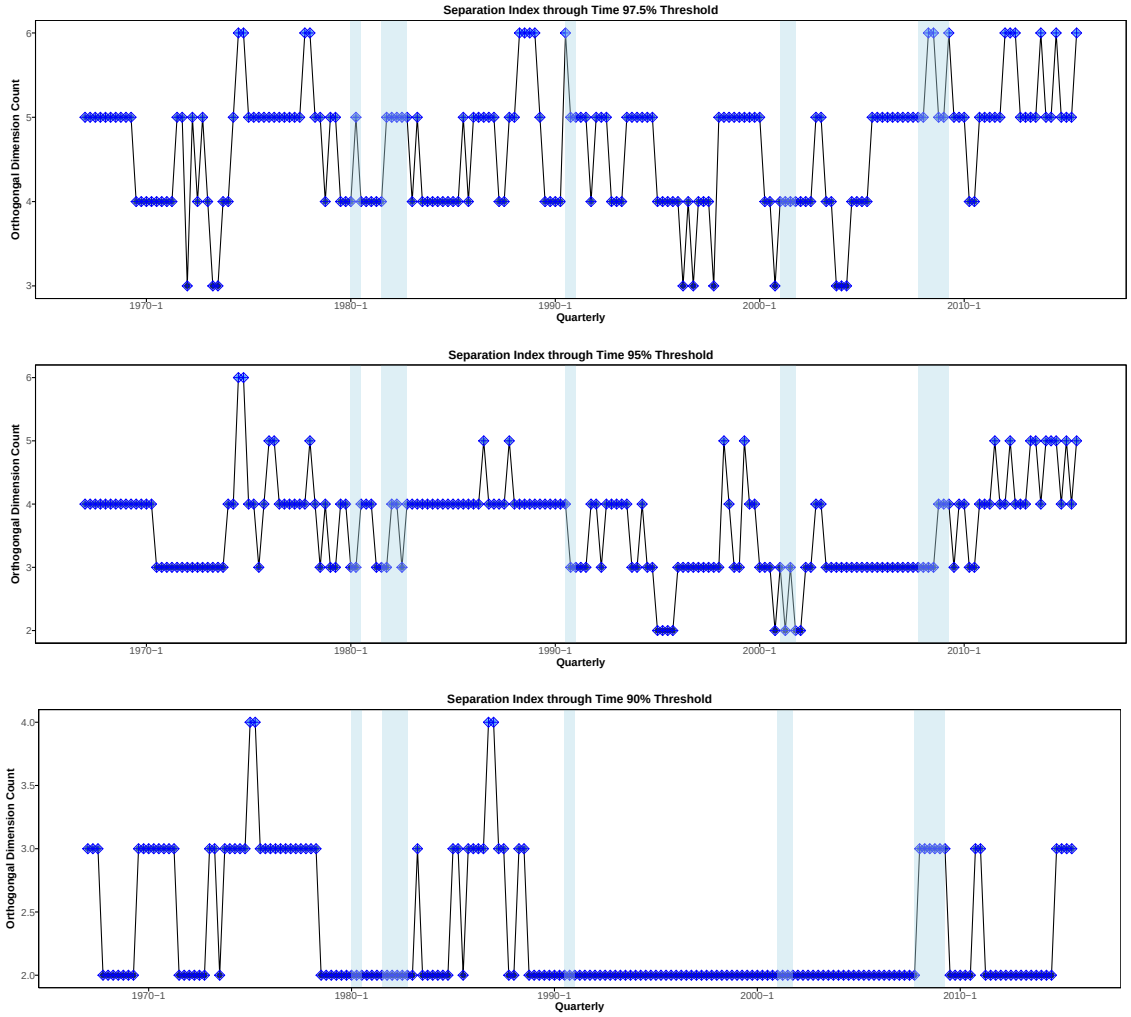
Using the full BTM 25 xsection, we compare the dimension of the representation of 97.5% variation threshold we find the inclusion of the nominal forecasts increases the dimension between one and three degrees. Heuristically, the dimensions can be thought of as state variables.

The fact that forecasts improve for horizons inside of four years as we bring the window T_m down from $T_m = 15$ to $T_m = 12$ and $T_m = 8$ is not surprising if the filtering procedure is working as expected. These choices represent an increase in the mixing time holding a threshold fixed. As a result, higher frequency fluctuations are emphasized at the expense of lower frequency transitory variation.

1.5.7 Summary of Empirical Method

To summarize, pick a threshold $\epsilon_0 > 0$ for the mixing times $N_t(\epsilon_0)$ and target $N_0(\epsilon_0)$. This choice implies some sample window size T_m such that the mixing time evaluated on the window $\tau_m(t)$ is near $N_0(\epsilon_0)$. Within each sample window $\tau_m(t)$, nominal forecasts are fitted for each of the risk factor returns and combined with the contemporaneous realized returns to create an augmented panel. Calculate a singular value decomposition (SVD) of the series in this panel and extract the significant dimensions. Transform the diagonal elements using $\Lambda \mapsto C(T_m)(1-\Lambda)^{-1}$. Forecast the component dimensions $t \to t+1$ by linear autoregressions

Figure 1.5: Orthogonal Dimensions Count 85% – 97.5% -thresholds



(a) The time series of the number of orthogonal dimensions needed to explain each threshold percentage of the variation from that date on a rolling historical 15-year window. Quarterly Fama -French 3-factor and Carhart model returns data are from 1967 $Q1$ to 2016 $Q4$. NBER recessions are in blue.

(do not include the first component in general). Regress the target portfolio returns on the components in window $[T_m - t, t]$ and keep the coefficient estimates. The latent component forecasts give forecasts for the target portfolios by weighting the component forecasts with the portfolio's coefficient estimates.

The specialized case where the best predictor is the spectral gap follows from equilibrium asset pricing theory, as described in the modeling sections. However, statistics contained in the trailing solutions to the eigenvalue problems are myriad. The above described predictive routine applies generally, as long as a first-order transition representation is justified.

1.6 Conclusion

We extract predictable components from priced risk factors and show these components can be used to improve allocation efficiency in real time. Latent, transitory components of factor risk prices contain valuable information about near and medium term evolution of the state of the economy. Novel evidence connecting time-series predictability and time-varying risk prices for common factors in equity markets is provided.

The *spectral gap* measures the fraction of volatility concentrated on long-run shocks, varies through time and predicts market returns. We use these factors to predict the market as well as several portfolios including value, size and momentum. For the market and value, we find out-of-sample R^2's of 10.8% and 14.7% respectively, for annual returns. For size, annual returns are 13.7% predictable out of sample. Momentum predictability is low at the short end, but reaches nearly 30% at the three year horizon. More generally, we document a heterogeneous term structure of predictability across types of portfolios. Most strikingly, the value premium predictability is concave and most predictable at the one-year horizon, while momentum predictability is convex. This finding contributes an interesting wrinkle to the value and momentum "puzzle" regarding the high average returns but significant negative

correlations of these series.

The discrepancy between the variation of the permanent and transitory components is a well studied object that emerges in a range of contexts in nature. Because the general representation for the concentration of volatility on the leading factor is measured by (one-minus) the difference between the first two eigenvalues of the Laplacian of a dynamical system, it is called the *spectral gap*. In asset markets, the spectral gap and related statistics are found by specializing a known decomposition of Markov transition dynamics. This decomposition identifies a factor structure of expected returns from the permanent and transitory fluctuations. The *common* sources of transitory fluctuations comprise the expected return *factors*. Naturally, the expected return factors are also the common sources of time-series predictability.

We begin with priced risk factors so we know the benchmark portfolios are priced, and find that incremental efficiency gains associated with identifying the expected return factors are also priced. Incremental gains from timing HML are compensation for exposure to business - cycle news, while timing Momentum compensates for permanent wealth shocks. Timing gains are not priced for size. We isolate the component of market returns that are priced but find a residual process with significant variation that has no cross-sectional pricing power.

Our findings suggest investors are able to allocate capital with more precision using past returns data alone than previously indicated. This distinction is most relevant in economies where we distinguish between revisions in state variables given a known, fixed model and economies where dynamical features require shocks to the model. For example, it is not only the case that the evolution of technology is uncertain and risky, but also the case that the way markets equilibrate is uncertain, particularly in the context of the given technological uncertainty. Because this problem immediately becomes high dimensional, the most natural device for managing it is high dimensional random variables in the form of random local transition dynamics. This modeling choice is simplified in practice because the distribution

of the eigenvalues (and eigenfunctions) of the system are well studied and tractably summarize high dimensional information.

Our characterization of the latent components has implications for economic modeling. Rational agents making allocations optimally in an asset pricing model must be endowed with the spectral data. In models where data are initialized at a deterministic invariant distribution this distinction is immaterial. In more general models, incorporation of spectral data into the decision makers' information set can impact model implications.

The predictable component in market returns is related to slow moving cash-flow yields, placing restrictions on parametrizations of stochastic discount factor models. Two natural theoretical benchmarks are i.) persistent level shocks to growth under representative Epstein Zin preferences, and ii.) i.i.d. growth (i.e., output is a random walk) under representative ambiguity aversion. Hansen (2011) points out that ambiguity averse investors ex-post look like rational expectations investors if the equilibrium data generating process is the worst-case model. For plausibly indistinguishable models, the worst case model is the long-run risk model. The findings and methods in this paper suggest an opportunity to distinguish these stories in finite samples by looking directly at dynamics in the spectral data (after making some assumptions regarding the underlying equilibrium and its data generator). Given the factor structure of time-varying discount rates, surprise revisions to persistent premia are priced cross-sectionally (ex-post in finite samples) if the long run risk model is the true model, but not necessarily when the representative investor is ambiguity averse.

Our analysis revealed the importance of the countercyclical concentration of return volatility on permanent shocks - both as an economic concept and a convenient modeling object. This suggests a similar analysis can be fruitful in a variety of other asset markets. Because these analyses are cast in terms of objects that are common to many asset pricing models, new empirical evidence implies tractable restrictions to the set of plausible parametrizations of stochastic discount factor models.

Chapter 2

Banking with Risky Assets

Introduction

A significant fraction of the typical individual's net worth is risky but not insurable. This comes with direct costs, from uninsurable losses, but also indirect costs. Losses in the uninsurable component of wealth change the composition of an investor's portfolio. This in turn requires rebalancing through the remaining, *tradeable* component of wealth, at prevailing prices. Direct losses are amplified if market prices are low and selling risky claims is costly.

We model a financial intermediary designed to mitigate these costs. The intermediary holds the tradeable component of investors' wealth on it's balance sheet and issues demandable claims as remuneration. Demandability allows bank investors to withdraw up to the cash value of their deposited endowment in exchange for a pro-rata share reduction in the risky asset. This arrangement improves welfare when markets are incomplete. The intermediary partially insures shocks directly by facilitating a transfer proportional to one minus the net capital gain on traded assets. The risk-sharing mechanism links the intermediary's balance sheet and the equilibrium pricing kernel. This link provides novel testable implications.

Investors can rebalance implicitly through the bank rather than directly in securities

markets when bank assets are *risky*. To illustrate how this works, note that shocks to the uninsurable component of wealth are independent of the market, and that ex-ante, each investor holds her optimal portfolio. Now, ex-post low private income investors have an unexpectedly high fraction of their total wealth in the risky claim. In response, they liquidate risky holdings by exercising the cash option. Ex-post high private income investors are conversely underexposed to the aggregate claim - but, they acquire a leveraged position in the risky asset *passively*. In the simplest case, the net result is that each type attains their optimal portfolio exactly.

Because the withdrawals are made ex-post based on portfolio demand, state dependence of the bank's capital structure reflects a wealth-weighted average of investor's effective risk aversion. The ex-post rollover rates capture the representative risk prices in each aggregate state. Moreover, the ex-ante asset levels capture the representative inter-temporal marginal rate of substitution (IES). As a result, our theory ties the bank balance sheet to a complete description of the equilibrium pricing kernel.

A challenge for theories linking intermediation to asset prices through individual preferences is that intermediary and market institutions should be jointly accessible. In the context of explaining an institution's role in resolving incentive conflicts, a common assumption is to prohibit investors from accessing markets directly. Institutional preferences that reflect the modified interests of a collection of individuals can then stand-in for a marginal investor. However, the market restriction is empirically implausible.[1] Resulting explanations for equilibrium asset prices are tenuous. A theory linking banks to asset prices through individual preferences cannot rely on restricting access of individuals to markets.[2]

[1]For exchange-traded equities and indices, the assumption is implausible. The importance of the incremental effort required to access options markets, bond markets, REITs etc, is debatable. In the latter case, lower observed participation rates are not because of inability to access, but rather a choice not to access.

[2]In full nuance, the theory cannot produce a representative agent that is restricted. Subtler forms of heterogeneity in place of blanket restrictions, e.g., the Lucas family device (Lucas 1990), can produce a plausible theory. Lucas (1990) models a representative "family" by restricting family members to certain tasks within each period, but then aggregating decisions at the family level. We do not require this device.

Alternatively, by abstracting from preferences in the population, the institution can be endowed directly with preferences or interests and used as the representative agent.[3] However, this abstraction becomes costly in the face of welfare, benchmarking and policy analyses.[4] In contrast, a theory linking intermediation to the preferences of an investing population can be productively integrated in the normative space.

The theory presented in this paper links the intermediary balance sheet to investor preferences, and allows investors access to exchange and banking institutions simultaneously. This trade-off in each period is what keeps a positive measure of investors positioned in the bank. Moreover, the rebalancing motive is risk-based, while the deposit motive is IES-based. Together, they address the problem of time-consistent policies. Investors' ex-post policies are optimal solutions to the contemporaneous portfolio-choice problem. Thus, the mechanism addresses the enforcement problem in incomplete markets in addition to the verification problem.

Haubrich and King (1990) critique the view that banks uniquely produce liquidation options the are credible because of fragility. They argue that the bank services can be broken into a liquidity component, that can be provided in ex-post securities markets, and an insurance component, that can be replicated by a mutual fund with the right configuration of coupon payments and share purchases. However, our setting relies on aggregate risk for asset pricing. The securities markets do not provide liquidity when market price depend on the aggregate state. Although their original analysis is not done with aggregate risk, in principle a mutual could announce any coupon and share purchase policy made contingent on the aggregate state and thus may be able to reproduce the bank system allocations. We show that replication by a mutual is in fact not possible.

The key mechanism in our theory precluding replication by a mutual or other market

[3]Krishnamurthy (2014) models the intermediary's marginal value of reputation.

[4]Welfare ideally incorporates the effects on individual utility and efficiency including endogenous equilibrium effects, ruled out by this abstraction. Benchmarking against competing or complementary theories based on individual optimization is also limited.

institution is the *synthetic* leverage generated by the bank to accommodate the various claimants to its assets. When a bank financier makes a withdrawal, the bank debits the residual capital account. Bank capital becomes leveraged and the corresponding changes in bank capital risk are *synthetic*, because the bank does not need to clear its shares and liabilities in the market contemporaneously. In contrast, the mutual marks-to-market, in that changes in its liabilities must correspond to changes in its assets. Both the mutual and the economy are unlevered in equilibrium. The bank precludes contemporaneous unwinding of synthesized leverage, and generates both the concentrated risk and the negative cash position high types need to be indifferent to rolling over the bank position.

2.0.1 Related Work

This paper is motivated by at least two areas of research. The first is the theoretical literature on bank liability design based on Diamond and Dybvig (1983). This literature constitutes the basis for understanding endogenous intermediation liquidity creation. The second is the literature on asset pricing and financial intermediation, beginning with empirical work by Adrian, Etula and Muir (2015). Drawing on methodologies and outstanding questions from each of these areas, we show portfolio choice motives are sufficient to microfound a financial intermediary. Because the intermediary microfounded in this way is financed dynamically based on preferences for risk, the intermediary balance sheet is linked to the incomplete markets stochastic discount factor (SDF).

Diamond and Dybvig (1983) model a bank deposit contract that solves the ex-ante allocation problem for a large economy of individuals who are uncertain about the timing of their consumption needs, and where capital is only productive in the long term. Ex-post population frequencies of near and long term consumption are known and there is no aggregate risk. Deposits allow investors to delay their commitment to an allocation between short and long term investments until their consumption timing preference is revealed. Knowledge of

the ex-post population frequencies permits the bank to allocate the deposits between short and long term investments more efficiently ex-ante.

Haubrich and King (1990) argue that ex-post securities markets can provide the same liquidity as deposits when there is no aggregate risk. Moreover, they argue the bank per-se is not preferred over a mutual unless transactions in securities markets between individuals are restricted. We show that the ex-post securities market does not provide this liquidity when there is aggregate risk and prices vary across states ex-post. As a result, preference for the intermediary does not rely on restricted access to markets for aggregate claims. In fact, there is no need for the bank to price discriminate based on timing, because the trade-off between ex-post market prices and the common initial price of the bank claim ensures investors adjust ex-post funding predictably.

Using constant elasticity of substitution (CES) utility, Haubrich and King (1990) argue deposit contracts driven by consumption timing, such Diamond and Dybvig (1983), are driven by the inter-temporal elasticity of substitution (IES). Our theory extends the work of earlier theories to portfolio allocation motives. We show our ex-post withdrawal policies are set from a portfolio rebalancing motive rather than a consumption-savings motive. Ex-ante, the IES drives savings policy and impacts bank deposit levels, but ex-post, the withdrawal policy is a function of risk aversion only. Through this channel, the bank's capital structure is connected to risk-based asset pricing, and hence the equilibrium SDF.

The Diamond and Dybvig (1983) model and its progeny use a sequential timing protocol for deposit withdraws to show coexistence of inefficient bank-run equilibria. Multiple equilibria led to an insight about the fragility of a bank funding structure as the source of its strength. Depositors *en-masse* credibly threaten runs simply by owning the demandability option, thus providing a source of discipline for banks. If a bank makes a promise ex-ante to honor ex-post withdrawals, the threat of runs prevents the banks from reneging on their promise. Common knowledge of this device ex-ante makes formation of the bank possible.

We do not use a sequential service constraint. However, the rebalancing motive gives

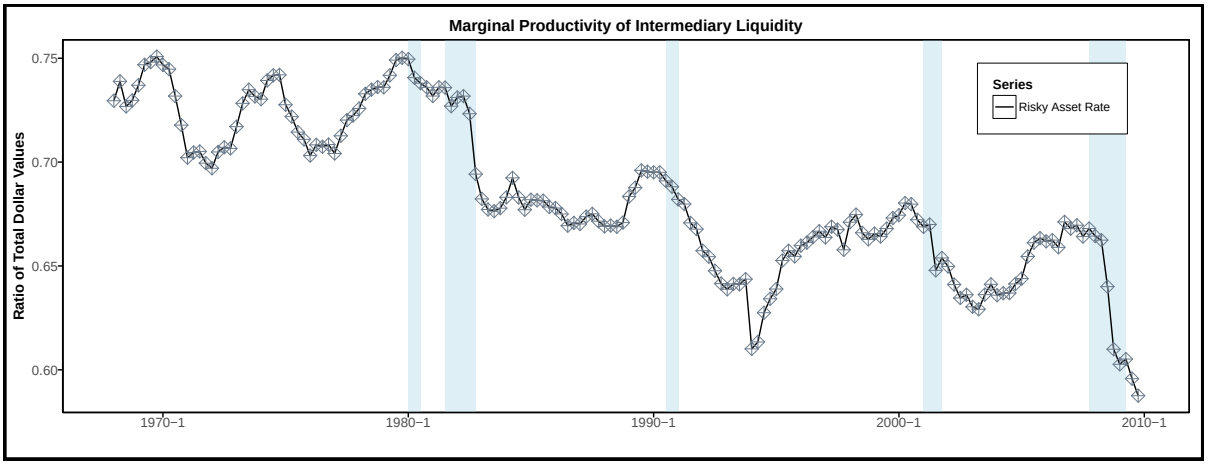
novel insight into bank funding trade-offs. Risky assets that have near-zero cash flows with positive probability preclude bank formation. Disaster risk impacts the ability of the bank to credibly produce liquidity, requiring the bank hold a cash buffer. In contrast, risky assets in the absence of disaster risk can improve the efficiency of liquidity production. While runs as sunspot equilibria are not a feature of our model, we show bank funding is contingent on the competitiveness of its expected return.

Our study has implications for the coexistence of intermediation and securities markets with unrestricted access. Allen and Gale (2010) study an economy with aggregate risk where investors have restricted access to securities markets, and derive the constrained-optimal asset holdings for a financial intermediary. We find that when intermediation appeals to portfolio motives, investors trade-off bank financing with direct trade in securities markets based on the direction of their trade and market prices. Expected utility is maximized when both institutions are accessible.

Our investigation is influenced by the recent literature connecting intermediary balance sheet dynamics and asset prices. A key empirical contribution is Adrian, Etula and Muir (2015), who find that asset return exposure to shocks to dealer leverage can explain cross-sectional variation in average returns. He, Kelly and Manilla (2017) find that a measure of bank capital can be used to explain the cross-sections of asset classes outside of equity and bond markets. Adrian et al (2012) shows a related measure of leverage has predictive power for market returns.

The empirical literature also studies the balance sheet dynamics of financial institutions. Adrian et al (2010), and Boyarchenko et al (2011), document leverage dynamics of broker dealers and commercial banks respectively, each of which are liquidity producers. Krishnamurhty et al (2014) document that commercial banks take on debt to acquire risky assets in bad times, which are being sold off by non-liquidity producing investment institutions like hedge funds, mutual funds, pensions and others. In Fig.1, we plot the rate of high-risk assets to liquid liabilities. We see a sharp downturn in every recession, as well as a low frequency

Figure 2.1: Ratio of High-Risk Assets to Liquid Liabilities



(a) The rate at which the highest risk assets contribute to liability-side liquidity production drops sharply in recessions. Separately, the rate exhibits a secular trend downward. Quarterly balance sheet data from 1967 $Q1$ to 2012 $Q4$ are from the Flow of Funds, Board of Governors of the Federal Reserve. We use private depository institutions, issuers of asset-backed securities, and securities brokers and dealers to measure liquidity production. The ratio of high risk assets to liquid liabilities is calculated by classifying liquid liabilities as large time deposits, uninsured checkable and savings deposits, asset backed commercial paper (ABCP) and repurchase agreements. Risky assets are corporate equities, mutual fund shares, and private residential and commercial mortgage-backed securities (MBS). NBER recessions are in blue.

trend downwards.

A theoretical literature in this area includes and Krishnamurthy (2013), Brunnermeier and Sannikov (2015) and Adrian and Boyarchenko (2015). Our theory differs in emphasis and methodology, and as such the two approaches are complementary. IAP theories do not aim to justify the intermediary in equilibrium. These theories extend dynamic asset pricing models designed to produce quantitative statements about the dynamics of asset returns in a variety of experimental settings to cases where a financial intermediary is marginal in securities markets. They also provide economic insight into how financial intermediation impacts risk and return.

In our model economy the financial intermediary is endogenous. The asset pricing implications in our model arise because of the connection between a preference for holding risky assets in the cross-section and the aggregate bank capital structure. This connection is an equilibrium outcome. The cost of this insight is that quantitative exercises in our stylized model are difficult to justify. However, qualitative predictions from our model generate

testable implications that are valid in dynamic settings. We provide reduced-form empirical evidence that supports our theory.

2.0.2 Example

To highlight the economic mechanism we present an example. The rigorous description of the model begins in section 1.

The Setting Consider two investors, I_1 and I_2, who each own equal claims $\frac{1}{2}V_0$ to a project valued V_0. The project pays an uncertain amount Y two periods from today. Each investor i is also entitled to an uncertain cash payment n_i that cannot be insured. The investors have log utility over final wealth. For simplicity we assume $n_i \in \{-\Delta, \Delta\}$ and $n_1 + n_2 = 0$. Each of the two configurations occurs with equal probability. The investors begin with equal stores of "cash" $\frac{1}{2}C > 0$ and have access to a storage technology with per-period gross return normalized to one. Total initial wealth is $W_0 = V_0 + C$.

Timing In the first period $t = 1$, the payments n_i are revealed. Prospects for the project payout Y are also revealed through a signal $m \in \{L, H\}$ corresponding to low and high productivity. In $t = 2$, the realization Y_m is either above (denoted "A") or below ("B") its conditional forecast $Y_{m,A} > E[Y_{m,k}|m]$, or $Y_{m,B} < E[Y_{m,k}|m]$. Claims to the project output are traded competitively in periods $t = 0, 1$. Time $t = 1$ prices are V_m.

Implications Because the investors are identical ex-ante, and the income n_i is uninsurable, there is no incentive to modify holdings at time-zero. Each investor's initial portfolio Π_0 can be written $\Pi_0 = (\alpha_0 W_0, (1 - \alpha_0)W_0)$ for initial wealth W_0. When normalized by wealth, the first entry corresponds to the fraction of wealth invested in equity at $t = 0$, and the second entry is the fraction of wealth held in cash. With no trade, $\alpha_0 = V_0/W_0$.

At $t = 1$, idiosyncratic income n_j is realized, along with the news about productivity

m. We let $i = 1$ correspond to the unexpectedly wealthy investor: $n_1 = \Delta > 0$. The two investors have identical shares of the risky asset at the *beginning* of the first period $a_0 = \frac{1}{2}$. However, idiosyncratic shocks produce heterogeneous wealth levels $W_j = a_0 V_m + \frac{1}{2}C + n_j$. As a result, the unlucky investor has an oversized rate of investment in the risky project $a_0 V_m / W_2 > a_0 V_m / W_1$.

Securities Trading The investors adjust positions until investment rates are equalized at the *end* of the first period, i.e., $\Pi_{1,1}/W_{1,1} = (\alpha_1, (1-\alpha_1)) = \Pi_{1,2}/W_{1,2}$. Subscripts (t, i) in $W_{t,i}, \Pi_{t,i}$ indicate the period and the investor, respectively. The market clearing share is $\alpha_1 = V_m \left[C + V_m\right]^{-1}$. Changes in individual wealth levels do not impact individual portfolio weights α. With log utility, each investor splits their idiosyncratic income pro-rata $n_i = [\alpha_1 n_i]_{\text{Equity}} + [(1-\alpha_1) n_i]_{\text{Cash}}$. Transactions are made at the ex-post market prices.

Written with terms gathered by position, the final wealth shares are

$$\begin{aligned} W_{1,m,1} &= V_m \left[\frac{1}{2} + \frac{\Delta}{W_{1,m}}\right] + C\left[\frac{1}{2} + \frac{\Delta}{W_{1,m}}\right] \\ W_{1,m,2} &= V_m \left[\frac{1}{2} - \frac{\Delta}{W_{1,m}}\right] + C\left[\frac{1}{2} - \frac{\Delta}{W_{1,m}}\right] \end{aligned} \tag{WS.0}$$

The first term, scaled by the share price V_m, is the market value of each investors equity position. The second term is the risk-free position. Anticipation of the ex-post distribution of net-worth in WS.0 is reflected in initial prices.

The Bank The investors instead create the following arrangement. The investors deposit their claims in a bank, and allow the bank to hold the risky claim as an asset. The investors now hold the liabilities of the bank instead of the claim to the project. In turn the bank includes provisions in the liabilities that allow the investors to withdraw any amount of the cash value of their deposit before the project matures, at which point the proceeds are paid to the residual claimants of the bank assets, and the bank is dissolved.

Implications with the Bank For simplicity, investors deposit the cash value $b_0 = \alpha_0 \Delta$ of their risky endowment, or equivalently, $\alpha_0 \frac{\Delta}{V_0}$ shares of their risky endowment, in the bank. This leaves a direct equity position with cash value $[\frac{1}{2} - \alpha_0 \frac{\Delta}{V_0}]V_0$, and the cash position $\frac{1}{2}C$. Write $\mathbf{b} = 2\Delta \frac{\alpha_0}{V_0}$ for the fraction of the risky claim intermediated at time-zero.

Now, at $t = 1$, in a recession $m = R$, the investor with $n_2 = -\Delta < 0$ withdraws cash from the bank in the amount of b_0. This transaction liquidates $\frac{\alpha_0}{V_0}\Delta$ shares with market value $\frac{\alpha_0}{V_0}\Delta V_m < \frac{\alpha_0}{V_0}\Delta$. The low type has no remaining exposure to the bank. We can calculate her shares explicitly: $\frac{b_0 - \alpha_0 \Delta}{2b_0 - \alpha_0 \Delta} = 0$. The high type passively acquires the residual bank position: $\frac{b_0}{2b_0 - \alpha_0 \Delta} = 1$. With no further action, the resulting portfolios are

$$\frac{\Pi_{1,1}}{W^b_{1,1}} = \underbrace{\left[\frac{1}{2} - \frac{\alpha_0 \Delta}{V_0}\right] V_m}_{\text{Direct Equity}} , \underbrace{\mathbf{b}\, V_m}_{\text{Bank Liabilities}} , \underbrace{\frac{1}{2}C + \Delta(1 - \alpha_0)}_{\text{Cash}}$$

$$\frac{\Pi_{1,2}}{W^b_{1,2}} = \underbrace{\left[\frac{1}{2} - \frac{\alpha_0 \Delta}{V_0}\right] V_m}_{\text{Direct Equity}} , \underbrace{0}_{\text{Bank Liabilities}} , \underbrace{\frac{1}{2}C - \Delta(1 - \alpha_0)}_{\text{Cash}}$$

By combining the bank and direct equity exposures into a single equity position, we can write the portfolios

$$\frac{\Pi_{1,j}}{W^b_{1,j}} = \alpha_1, (1 - \alpha_1) \qquad\qquad j = 1, 2$$

for wealth shares $W^b_{1,j} = W_1 \left(\frac{1}{2} + (-1)^{j-1}\Delta\right)$. Through the bank's balance sheet, the withdrawal policy of the unlucky investor successfully implements her required portfolio adjustment, *as well as the portfolio adjustments of the lucky investor.*

Remark Relative to the incomplete markets wealth share $W_{1,2} = \frac{1}{2}W_1 - \Delta$, the low-type saves $\Delta[V_0 - V_m] > 0$. Gains arise because the low income investor liquidates risky holdings at cost $\alpha_0 \Delta \frac{V_{1,R}}{V_0} < \alpha_0 \Delta$, thereby implementing an implicit share transfer from the ex-post high type. The high income investor is indifferent to this transfer at prevailing prices.

Discussion In this example, the converted payment is in the form of an I.O.U. in the amount of the withdrawal $\alpha_0\Delta$, to be paid when the output is realized. Residual claimants are entitled to the output *net* of the I.O.U., and the residual claimant's position is commensurately *more concentrated*. The change in exposure for investor I_1 is $0.5b_0[b_0]^{-1} \mapsto b_0[2b_0 - \alpha_0\Delta]^{-1} > 0.5b_0[b_0]^{-1}$. Passive rebalancing works when *bank assets are risky* by allowing exercised cash options to leverage residual exposures.

Policies in the $m = G$ Case In this example, the low-type can liquidate the risky claim at a better rate on the market directly when net capital gains are positive. The high type cannot acquire assets through the intermediary balance sheet unless the low-types exercise cash options. As a result, bank exposures remain symmetric, and rebalancing is carried out in securities markets.[5]

Put Option on Bank Assets The ex-post transfer $\Delta[V_0 - V_m]$ can be written ex-ante as a put option on the bank's assets

$$\tau = [K^* - S_1]^+ 1_{\{n_1 = -\Delta\}}\Delta$$

The liabilities embed an option that will only be exercised by investors with negative idiosyncratic shocks in bad times, when net capital gains are negative. Liquidity is created by allowing bank financiers to lock-in the ex-ante share price, through the strike $K = V_0$, as a contingency for the event that an ex-post liquidation is needed when market prices are low.

Remark For an ϵ-fee on deposits, investors will finance the bank by exactly the amount they will need to withdraw in the bad aggregate state. The reason is that, in the good state, the low type would prefer to liquidate risk claims on the market because the implied share

[5]In a more general setting, in good times, the ex-post high type can exercise a call option on bank assets by supplying new cash funding for the bank. This policy leaves the low-type indifferent given their exposure to the risky assets is diluted proportionally.

price of her bank deposit is lower.

2.0.3 Organization

Complete markets, incomplete markets and intermediated incomplete markets versions of the model are developed for contrast. Section 1 details the resources, participants and market arrangements that serve as a benchmark to each of the versions we develop. Section 1.4 specializes to the economy with financial intermediation and derives equilibrium allocations and prices. The complete and incomplete markets environments are specialized and solved in A.0. Section 2 details comparative implications. Proofs based on standard arguments are relegated to Appendix A. The final section 4 discusses a handful of applications, including policy implications.

2.1 The Model

2.1.1 Environment

There is a continuum $\mathcal{I} := [0, 1]$ of ex-ante identical investors. Investors have log utility over terminal wealth and are endowed with equal claims e_0 to terminal output Y. Y can take four possible values $Y \in \{Y_{GA}, Y_{GB}, Y_{RA}, Y_{RB}\} =: \mathcal{Y}$, where $Y_{GA} > Y_{RA}$ and $Y_{GB} > Y_{RB}$. Uncertainty is resolved over two periods. In the first period, a public signal indicates growth G or recession R. In the final period, realized output Y will will be above or below market expectations. For example, Y_{GB} corresponds to below expected output in the growth regime. Aggregate states are denoted s_m for $t = 1$ and $s_{m,k} \in S$ for $t = T$ with indices $m = R, G$; $k = A, B$. We write $Y(s_{m,k}) =: Y_{m,k} = Y$. Probabilities are mutually independent, $\Pr(s_{m,k}) = \pi_{m,k} = \pi_m \pi_k$, with marginals $\Pr(s_m) = \pi_m$ and $\Pr(s_{m,k}|m = R) + \Pr(s_{m,k}|m = G) = \pi_k$.

Each investor is also endowed with a nontradeable claim to income n_j distributed ac-

cording to

$$n_j = \begin{cases} \Delta_0 + \Delta & \text{with } \Pr(n_j = \Delta_0 + \Delta) = \frac{1}{2} \\ \Delta_0 - \Delta & \text{otherwise} \end{cases}$$

with $|\Delta| < \Delta_0$. The random variable n_j is revealed in the intermediate period. n_j is independent of the aggregate signal $m \in \{R, G\}$ and i.i.d. in the cross-section. Joint probabilities for $(s_{m,k}, n_j)$ are $\pi_{m,k,n_j} = \frac{1}{2}\pi_{m,k}$. There is no aggregate income risk. Total income is $Y_{m,k,0} := \Delta_0 + Y_{m,k}$ in every state. The resolution of uncertainty is depicted by a binomial tree in Figure 3.

An equity claim on output Y is traded at $t = 0$ and $t = 1$. Equity shares are fixed at one. Equity prices V_0 and $V_{1,m}$ are determined equilibrium. Time-zero share purchases in excess of the endowment are written a_0. After time-zero, share adjustments are denoted a_j. We write *gross* positions as a proportion of individual wealth α_0 and α_j. Allocations a, α are functions of time, individual wealth and the aggregate state.[6].

Finally, investors are endowed with equal deterministic amounts of a durable numeraire, "cash," written $\omega_0 > 0$. A riskless storage technology in infinitely elastic supply yields gross return normalized to $R = 1$. The securities markets open in response to news, as depicted in Figure 1.

Expenditures at time-zero are constrained by *tradeable* wealth $W_0 = \omega_0 + e_0$ where $e_0 = \mathbb{E}^Q[Y]$. The risk-neutral measure Q is determined in equilibrium. Investors can adjust their initial positions e_0 in the risky claim through choice of a_0 subject to $a_0 V_0 - e_0 \leq \omega_0$. Investors will forego trade if it is optimal.

[6]The shorthand $a_0 = a(0, \cdot, \cdot)$ emphasizes initially identical policies, while $a_j = a(1, n_j, m)$ emphasizes the type-j dependence of allocations made at $t = 1$ given signal m. The same applies to α_0, α_j.

2.1.2 Equilibrium

Sequences

Write the investor's initial endowments $\boldsymbol{q}_0 = (V_0, \omega_0)'$ and define $q_0 = \boldsymbol{q}_0' \cdot \mathbf{1}$. The condition $e_0 = V_0$ gives $q_0 = W_0$, the initial level of tradeable wealth. We summarize each investor's final cash position $\omega_{j,m} := \omega_0 - a_0 V_0 + n_{1,j} - a_j V_{1,m}$, and the corresponding equity position $A_j := (1 + a_0 + a_j)$. Write $\theta_{T,j}$ for the period-T gross return on one dollar invested at time-zero for ex-post type j. $\theta_{T,j}$ is cum-income. It includes the income in $t = 1$, n_j, and the associated gain or loss on that investment between $t = 1$ and $t = T$.

Individual optimization problems are

$$
\begin{aligned}
J(q_0; V_0) \;=\; & \max_{\boldsymbol{a}} \mathbb{E}\left[\log(\theta_{T,j})\right] && \text{(1.A)} \\
\text{s.t.} \quad & a_0 V_0 - V_0 \leq \omega_0 \\
& a_j V_{1,m} - (1 + a_0) V_{1,m} \leq (\omega_0 - a_0 V_0) + n_{1,j} \\
& \theta_{T,j} = A_j Y_{m,k} + \omega_{j,m}
\end{aligned}
$$

where $\boldsymbol{a} := (a_0, \{a_j\}_j)$. The second inequality is reproduced for every $(n_{1,j}, m) \in \{n_{1,1}, n_{1,2}\} \times \{R, G\}$ and the final equality for each $k \in \{D, U\} | (n_{i,j}, m)$. Expectations are with respect to the joint distribution of productivity, prices and income $(\mathcal{Y}, V, \{n_j\}_j)$.

The recursive analogue 1.B to the objective 1.A is provided in section 6.0.1.

Normalization We normalize $W_0 \equiv 1$ without loss of generality. Investors with log utility over final wealth care only about single-period gross returns. Aggregate wealth is not a state variable, although market incompleteness requires that individual wealth $q_{t,j} = q_{t,j} W_0^{-1}$ is a state variable for every individual. Standard arguments based on homothetic preferences, given in section 6.0.1 of Appendix A, justify this choice of state vector.

Resources

Put $\overline{Y}_{m,k} := Y_{m,k,0} + \omega_0$. In what follows, we distinguish between expectations over idiosyncratic income states and ex-post aggregation across population types by writing $\mathbb{E}_{n_j}$ and $\sum_j \pi_j$, respectively.

Equilibrium The equilibrium is determined when every trader optimizes 1.A or 1.B and the following markets clear:

$$\begin{aligned} &\sum_j a_{0,j} = 0 && && (1.C)\\ &\frac{1}{2}\sum_j \alpha_j \tilde{W}_{1,j} = V_{1,m} && m \in \{R, G\} \\ &\frac{1}{2}\sum_j \theta_{2,j} = \overline{Y}_{m,k} && k \in \{U, D\}|m \end{aligned}$$

where we have applied $\pi_j = \frac{1}{2}$. The first two lines ensure securities markets clear at times $t = 0, 1$. The first line is simply $a_0 = 0$, or equivalently, $e_0 = V_0$, implying $W_0 = \omega_0 + V_0$. The second line states total shares are fixed at $1 = \frac{1}{2}\sum_j \overline{a}_j = \frac{1}{2}\sum_j \frac{\alpha_j W_{1,j}}{V_{1,m}}$. The third line is the final accounting for resources in terms of goods supply $Y_{m,k}$ and cash ω_0. When $k \in \{A, B\}$ is realized, output $Y_{m,k}$ is distributed according to the equity holdings A_j. Dependence on m is suppressed in some notation, e.g., $W_{1,j} = W_{1,j,m}$.

2.1.3 Discussion

Heterogeneous Wealth To highlight the rebalancing policies $a_j = a_j(n_{1,j}, m)$, we write the beginning of period wealth $W_{1,j}^-$ for each ex-post type j,

$$\begin{aligned} W_{1,j}^- = &\underbrace{W_0\,\alpha_0\,\frac{V_{1,m}}{V_0}}_{\text{Equity Capital Gain}} + \underbrace{W_0\,(1-\alpha_0)}_{\text{Storage}} + \underbrace{n_{1,j}}_{\text{Non-tradeable Income}} \\ = &\; W_1 + n_{1,j} \end{aligned}$$

In the first line, the first two terms are identical for every investor, suggesting we take $R_0 := R_{0,j} - n_{1,j}$ and write $W_1 = W_0 R_0$, giving the second line.[7] Now, write the end of period wealth $W^+_{1,j}$, for each ex-post type j

$$W^+_{1,j} = \underbrace{a_j V_{1,m} + (1 + a_0)\, V_{1,m}}_{\text{Equity Position}} + \underbrace{\omega_0 - a_0 V_0 + n_{1,j} - a_j V_{1,m}}_{\text{Cash Position}}$$

$W^+_{1,j}$ captures the composition of the investor's *outgoing* portfolio, expressed in terms of share policies a_j.[8] The first term is the value of their equity position after rebalancing at market prices $V_{1,m}$, and the second term is the value of the cash position. [9]

Securities Markets Take $j : n_j = \Delta > 0$ and $i \neq j : n_i = -\Delta < 0$. Clearly $a^*_j > 0 > a^*_i$. From the incoming portfolios $W^-_{1,j} > W^-_{1,i}$, together with the identical initial positions a_0, we see that $a_0 V_{1,m}/W^-_{1,j} < a_0 V_{1,m}/W^-_{1,i}$. Type j has a relatively lower *incoming* equity investment *rate* than type i. Since $W^+_{1,h} = W^-_{1,h}$ for any h, rebalancing operates entirely through a_h. With log preferences, the policies $a^*_j > 0 > a^*_i$ are chosen to equalize the equity investment rates $\alpha_j = \alpha_i$ across types.

Benchmark Implications

In Appendix A.0, we solve the model with log utility the for complete and incomplete markets cases and detail the implications. Complete markets naturally produce a degenerate wealth distribution and a representative agent pricing kernel. Incomplete markets imply the ex-post wealth distribution is bimodal with support that varies with the aggregate state.

[7] $\alpha_0 W_0 = (1 + a_0) V_0$ relates the fraction of equity a_0 to the fraction of wealth invested in equity α_0. Therefore $W_0(1 - \alpha_0) = \omega_0 - a_0 V_0$ is the total cash position. Policies with subscript 0 are identical across investors.

[8] Policies $a_0 = \alpha_0 W_0 V_0^{-1} - 1 > 0$ correspond to an increase in the investor's risky position at time zero, $a_0 < 0$ represents a decrease, and $a_0 = 0$ gives the time-zero no-trade allocation.

[9] Recall that policies $a_j > 0$ represent an increase in equity levels while policies $a_j < 0$ indicate a decrease.

The corresponding pricing kernel takes the form of the complete markets kernel scaled multiplicatively by a term accounting for the wealth distribution. Details of the incomplete markets asset prices are reproduced alongside the intermediated-economy asset prices, in section 2.2. The intermediary economy is formalized in section 1.4.

2.1.4 Intermediated Markets

Technology

The bank allows investors to deposit a fraction of their tradeable endowment e_0, denoted in levels by $b_0 \leq e_0$, in exchange for a claim to bank assets. The claim embeds the option to convert any amount $k_j = k(n_j, b_0)$, up to the cash value of the deposit $k_j \leq b_0$, into a certain payment. Total bank financing aggregates b_0 over investors and is written $\mathbf{b}$. The financing level $\mathbf{b}$ acquires $\frac{\mathbf{b}}{V_0}$ shares of the risky claim for intermediation. The remaining shares $\frac{1}{V_0}(V_0 - \mathbf{b})$ are held directly by investors. In each period, investors can access both the bank and the market directly. At time-zero, bank investors bear identical exposures to the bank asset risk.

Ex-post, an individual financier i has the option to respond to n_i via the rollover policy $k(n_i, b_0)$. However, the rollover policies *en-masse* $\{k_j\}_{j \in \mathcal{I}}$ control the risk composition of the residual claim to bank assets. Thus, an optimal funding policy can only be evaluated given an assessment of the aggregate effect of all funding policies on the residual capital.

Bank Capital For total withdrawals $\boldsymbol{\kappa} := \sum_j k_j \pi_j$, an infinitesimal investor j_0 choosing $k_{j_0} \in [0, b_0]$ acquires the residual exposure

$$\mathbf{k}^+(k_{j_0}, \boldsymbol{\kappa}) = \frac{b_0 - k_{j_0}}{\mathbf{b} - \boldsymbol{\kappa}} \tag{K.1}$$

The quantity $\mathbf{k}^+$ reports the fraction of total claims to bank assets. An investment of one dollar at time zero corresponds to a $\mathbf{b}^{-1}$ ownership stake in the bank. Ex-post, with

no withdrawal by the dollar investor and an economy-wide withdrawal of $\boldsymbol{\kappa}$, the dollar investment corresponds to a $[\mathbf{b}-\boldsymbol{\kappa}]^{-1} > \mathbf{b}^{-1}$ ownership stake. For initial investment b_0 and ex-post policy k_{j_0}, we obtain K.1. From the stake $\mathbf{k}^+$, we can calculate its market value at time $t=1$, $\mathbf{k}^+(k_{j_0},\boldsymbol{\kappa})\, V_{1,m}\, \frac{\mathbf{b}}{V_0}$ and the corresponding ownership stake in equity $\mathbf{k}^+(k_{j_0},\boldsymbol{\kappa})\, \frac{\mathbf{b}}{V_0}$.

Payments to residual claimants are net of the bank's obligatory payments $\boldsymbol{\kappa}$, which are also deducted pro-rata. We define the dividend paid to a marginal unit of bank capital $D_k := [\frac{\mathbf{b}}{V_0} Y_{m,k} - \boldsymbol{\kappa}]$. Naturally, the residual position $\mathbf{k}^+$ entitles an owner to the cash flows $\mathbf{k}^+\, D_k$ in each final state $k \in \{B, A\}$.[10]

An individual financier i's rollover policy k_i therefore depends on the ex-post bank capital structure in addition to the income realization n_i. We occasionally write $k_j = k(n_j, b_0, \boldsymbol{\kappa})$ to emphasize this dependence. Optimal policies and the ex-post population frequencies $\pi_j\, j = 1, 2$ are common knowledge. An individual investor has no market power. Her optimal rollover policy takes the ex-post capital structure $(\boldsymbol{\kappa}, \mathbf{b})$ as given, where $\boldsymbol{\kappa} = \boldsymbol{\kappa}(m)$. The resolution of uncertainty is illustrated in Figure 2, along with the timing of allocation decisions.

Preferences

In the appendix A.7.1 we discuss a nonseparable preference specification for the banking model. A simple modification to the endowment to include a proportional dividend allows us to define preferences over consumption streams, which is a more natural specification for recursive utility. Below, we continue our analysis with logarithmic preferences.

[10] A k without index or argument always refers to final stage uncertainty $k \in \{B, A\}$, while indexed or functional $k_j = k(n_j, b_0, \boldsymbol{\kappa})$ are always contingent rollover policies.

Equilibrium

We invoke the result that $a_0 \equiv 0$. After incorporating the bank technology, every investor's time-zero objective can be written

$$
\begin{aligned}
J(q_0; V_0) \;=\; & \max_{\boldsymbol{a}, b} \mathbb{E}\left[\log(\theta_{T,j})\right] \qquad \text{(B.1)} \\
\text{s.t.} \quad & b_0 \leq V_0 \\
& a_j V_{1,m} - \left[1 - \frac{b_0}{V_0} + \mathbf{k}^+(k_j, \boldsymbol{\kappa}) \frac{\mathbf{b}}{V_0}\right] V_{1,m} \leq \omega_0 + n_j + k(n_j, b_0, \boldsymbol{\kappa}) \\
& k(n_j, b_0, \boldsymbol{\kappa}) \;\leq\; b_0 \\
& \theta_{T,j} = A_{j,m}(\mathbf{b})\, Y_{m,k} + \mathbf{k}^+(k_j, \boldsymbol{\kappa})\, D_k + \omega_{j,m} + k_j
\end{aligned}
$$

where $A_{j,m}(\mathbf{b}) := [1 - \frac{\mathbf{b}}{V_0} + a_j]$ is the fraction of equity held directly by investor j.

Resources In the banking economy, market clearing conditions 1.C interact with the definitions of b_0, $\mathbf{b}$ and k_j because ownership of the equity claim is partially intermediated.[11] We restate the market clearing conditions below and discuss the role of bank variables in equilibration. Having already imposed $a_0 \equiv 0$[12], we can write

$$
\begin{aligned}
& \sum_j a_j \;=\; 0 \qquad \text{(2.C)} \\
& \frac{1}{2} \sum_j \theta_{2,j} = \overline{Y}_{m,k} \qquad k|m
\end{aligned}
$$

[11]The definitions of $\{b_0, \mathbf{b}, k_j, \boldsymbol{\kappa}\}$ along with market clearing criteria from the incomplete markets model, 1.C, are alone sufficient for a well-defined equilibrium.

[12]The equivalent conditions for wealth rates α_t attain via the obvious substitution $a_j = \alpha_t W_{t,j}/V_{t,m} - 1$.

The first term in 2.C says net time $t = 1$ modifications through the market a_j are zero. A key definition is worth restating

$$\frac{1}{[\mathbf{b} - \boldsymbol{\kappa}]} \frac{1}{2} \sum_j [b_0 - k_j] = 1 \tag{Re.1}$$

The resource constraint $\sum_j a_j = 0$, together with Re.1, is equivalent to enforcing that equity is in fixed supply with shares normalized to one. Re.1 also ensures the total output paid to the bank is $\frac{\mathbf{b}}{V_0} Y_{m,k}$, which, together with the share accounting for direct holdings $(1 - \frac{\mathbf{b}}{V_0} + \frac{1}{2} \sum_j a_j)$, ensures total output distributed is $Y_{m,k}$. Finally, from Re.1 and the definition of D_k, the total level of precedent payments owed by bank capital owners is $\boldsymbol{\kappa}$.

The second term in 2.C is an accounting of final payouts made to individuals. It is identical to the third term in 1.C with the exception that $\theta_{2,j}$, given in B.1, is a function of bank policies $k_j, \{k_{-j}\}_{-j \in \mathcal{J}}$. The first term in the original clearing list1.C is subsumed by the fact that $a_0 \equiv 0$ and the definition $\mathbf{b} = \frac{1}{2} \sum_j b_0$.

Equilibrium with Intermediation *An equilibrium with intermediary financing is a set of allocation policies a_j, b_0, k_j and prices $V_0, V_{t,m}$ such that every investor optimizes B.1, markets clear according to 2.C, and the policies $k(n_j, b_0)$ and ex-post population frequencies π_1, π_2 are common knowledge.*

2.2 Implications

We develop the portfolio and bank funding policy implications of our theory, along with the corresponding asset pricing implications. Implications for industrial organization in the financial sector are postponed to section 2.4. We first state some key results.

Proposition 2.2.1 (Bank Financing Equilibrium) *An equilibrium in the economy with bank liability production exists and exhibits the following properties*

1. *Trade in the initial period organizes the bank, leading to welfare gains*

2. *Ex-post policies implement rebalancing through the bank's balance sheet when market prices are low* $k(n_1, \boldsymbol{\kappa}(s_R)) = 0$,$k(n_2, \boldsymbol{\kappa}(s_R)) = \alpha_0 \Delta$, *and through securities markets when market prices are high* $k(n_1, \boldsymbol{\kappa}(s_G)) = k(n_2, \boldsymbol{\kappa}(s_G)) = 0$

3. *Prices in the intermediated economy can be written in terms of the Lucas kernel*

$$\mathbb{M}[(s_{m,k})]_{IAP} = \left[M[\overline{Y}_{m,k}]_{Lucas}\right] e^{-s_{m,k}\eta_s - s_0 \bar{s}_{m,k}\zeta_s}$$

4. *Relative to the incomplete markets benchmark, the distribution of subjective valuations is more dispersed but with lower mean, and the distribution of wealth is less dispersed*

2.2.1 Policies

Heterogeneous Wealth We revisit the wealth expressions from section 1.3 in the context of the banking economy. With incomplete markets, incoming positions are symmetric up to the shocks n_j. In the banking economy, the positions $b_0 = b_0(V_0)$ nest a put option on risky assets that separates ex-post investors by type in bad times. When exercised, the options induce portfolio heterogeneity by implementing the ex-post swap of cash for risky claims at prices set ex-ante.

We can write the incoming and outgoing expressions from 1.3 in terms of intermediary positions b_0 and rollover policies k_j

$$W_{1,j}^{-} = \underbrace{\overbrace{[V_0 - b_0]\frac{V_{1,m}}{V_0}}^{\text{Market Position}} + \overbrace{b_0 \frac{V_{1,m}}{V_0}}^{\text{Bank Deposit}}}_{\text{Combined Capital Gain}} + \underbrace{\omega_0}_{\text{Cash}} + \underbrace{n_{1,j}}_{\text{Non-tradeable Income}} \tag{W.B.1}$$

$$W_{1,j}^{+} = \underbrace{[a_j + 1]\, V_{1,m}}_{\text{Market Position}} + \underbrace{\frac{\mathbf{b}}{V_0}\, \mathbf{k}^{+}(k_j, \boldsymbol{\kappa})\, V_{1,m}}_{\text{Bank Capital}} + \underbrace{\omega_0 + n_{1,j} - a_j V_{1,m} + k_j - \boldsymbol{\kappa}\, \mathbf{k}^{+}}_{\text{Cash Position}}$$

where the residual claim is $\mathbf{k}^{+}(k_j, \boldsymbol{\kappa}) = \frac{b_0 - k_j}{\mathbf{b} - \boldsymbol{\kappa}}$ defined in K.1.

Rollover Policies For a withdrawal k_j, the investor rolls-over a fraction $b_0^{-1}[b_0 - k_j]$ of her initial stake in the bank. The level k_j augments the *Cash Position* of the investor's portfolio, illustrated in $W_{1,j}^+$ of W.B.1. The corresponding increase in exposure to asset risk is consolidated in the *Bank Capital* ledger from the same expression, and is written $\frac{b_0 - k_j}{\mathbf{b} - \boldsymbol{\kappa}} \frac{\mathbf{b}}{V_0}$, or equivalently, $\mathbf{k}^+ \frac{\mathbf{b}}{V_0}$. The residual claims to bank assets $\mathbf{k}^+(k_j, \boldsymbol{\kappa})$ are by definition the bank capital.[13] For any withdrawal policy short of full divestment $k_j < b_0$, a third portfolio implication accounts for the precedent payments $\boldsymbol{\kappa}$ nested in the bank dividend D_k. These payments are recorded as the final entry in the *Cash Position* ledger, written $-\boldsymbol{\kappa}\,\mathbf{k}^+$, as a liability corresponding to the rollover policy k_j. Each of the three portfolio components corresponding to k_j are illustrated in W.B.1 via $W_{1,j}^+$.

Remark Investors implement the optimal allocation policy $\alpha_j = \alpha_i$ through a combination of intermediary leverage and direct trade in securities markets. In equilibrium, $\alpha_{1,h} = \alpha_1 = V_{1,m}\,[V_{1,m} + C]^{-1}$ for each $h \in \{1, 2\}$ as in the incomplete markets economy. However, the corresponding price levels and wealth shares differ. The implementation of $\alpha_i = \alpha_j$ is ex-post less costly for high marginal utility types, and therefore ex-ante expected to be less costly for every investor.

2.2.2 Asset Prices

We present and discuss asset pricing implications from the incomplete markets economy and the intermediated economy. Proofs and background details are given in Appendix A.6.0.

Time-zero We can express the time-zero incomplete markets NC pricing kernel in terms of the expected wealth distribution. One-period asset prices in the complete and incomplete

[13] We use the term *bank capital* to refer to the market value of equity for the bank, consistent with the nomenclature in banking. In this model, equity is only defined implicitly ex-ante, but following the aggregate withdrawals, we unambiguously refer to the residual claims as bank capital.

markets economies are described by

$$M[\overline{Y}_m]_{\text{Lucas}} = [\nu_0]^{-1} W_{1,m}^{-1} \pi_m \tag{L.0}$$

$$\begin{aligned} M[\theta_{1,h}(s_m)]_{\text{NC}} &= [\nu_0]^{-1} [\theta_{1,j}^{-1} + \theta_{1,-j}^{-1}] \frac{1}{2} \pi_m \\ &= [\nu_0]^{-1} W_{1,m} [(W_{1,m} + \Delta)(W_{1,m} - \Delta)]^{-1} \pi_m \end{aligned} \tag{NC}$$

where $W_{1,m} = V_{1,m} + \omega_0$. The proof, given in section 6.2.4, uses the $t = 1$ marginal value of wealth $[\partial J_{1,j}]^{-1} = \theta_{1,j}$ for each j, that are consistent with backward induction from the final-period shares I.θ. The shares simultaneously obey the time-zero Euler equations.[14]

Oversaving The NC kernel includes a component correcting for the distribution of wealth that strictly raises state prices relative to the complete markets benchmark. The additional component vanishes as income transfers become negligible

$$[(W_{1,m} + \Delta)(W_{1,m} - \Delta)]^{-1} - W_{1,m}^{-2} \; > \; 0 \qquad \Delta > 0$$

$$\lim_{\Delta \searrow 0} W_{1,m} [(W_{1,m} + \Delta)(W_{1,m} - \Delta)]^{-1} = W_{1,m}^{-1}$$

NC reduces to L.0 with $\Delta \searrow 0$. NC accommodates higher demand for savings when some states of the world are uninsurable. [15,16]

Time-one Market prices at time $t = 1$ are complete markets prices, but with heterogeneous investors. With log utility, the representative and heterogeneous-investor pricing kernels are

[14]Time-consistency for incomplete markets, in the sense of Marcet and Marimon (1997),(2012), requires ex-post optimal policies agree with the policies that made ex-ante allocations optimal.

[15]Over-saving in incomplete markets with convex marginal utility is well studied, see e.g., Weil (1989) and Mankiw (1986).

[16]Demand for savings grows proportionally with the fraction of net-worth that is nontradeable. State prices rise in equilibrium because the market cost of postponing consumption must offset its increased demand.

equivalent

$$
\begin{aligned}
M[\overline{Y}_{m,k}|m]_{\text{Lucas}} &= [\partial J_1]^{-1}\, \overline{Y}_{m,k}^{-1} \pi_{m,k} && \text{(L.1)} \\
M[\theta_{2,j}(s_{m,k})]_{\text{Hetero}} &= \frac{\theta_{2,j}^{-1}}{\partial J_{1,j}} \pi_{m,k} = \frac{\theta_{2,-j}^{-1}}{\partial J_{1,-j}} \pi_{m,k} && \text{(H)} \\
&= \frac{1}{2}[\partial J_{1,j}^{-1} + \partial J_{1,-j}^{-1}] \overline{Y}_{m,k}^{-1} \pi_{m,k} \\
&= [\partial J_1]^{-1} \overline{Y}_{m,k}^{-1} \pi_{m,k}
\end{aligned}
$$

at time $t = 1$, for each $s_{m,k}$.

Log utility gives $[\partial J_{1,j}]^{-1} + [\partial J_{1,-j}]^{-1} = 2[\partial J_1]^{-1}$ with $[\partial J_{1,j}]^{-1} = W_{1,m} + n_{1,j}$ and $[\partial J_1]^{-1} = W_{1,m}$. The first line for M_{Hetero} equates valuations by type. The second line uses the wealth shares $\theta_{2,j}$ and represents the kernel by averaging $\partial J_{1,j}$ and $\theta_{2,j}^{-1}$ over types and then normalizing. The third line is true for any aggregation rule, e.g., averaging after normalization.[17]

Lucas Exchange Model The incomplete markets pricing kernel can be written in terms of the Lucas kernel and a multiplicative term reflecting imperfect risk sharing in the cross-section of the investing population. Write $\sqrt{\sigma_\Delta}(s_m) := \Delta / W_{1,m}(s_m)$. We express the incomplete kernel NC in terms of L.0:

$$
\begin{aligned}
M[\theta_{1,h}(s_m)]_{\text{NC}} &= \left[M[\overline{Y}_m]_{\text{Lucas}}\right]\ e^{-\log[(1+\Delta/W_{1,m})(1-\Delta/W_{1,m})](s_m)} \\
&= \left[M[\overline{Y}_m]_{\text{Lucas}}\right]\ e^{\sigma_\Delta(s_m)(1+\frac{1}{2}\sigma_\Delta(s_m))-o(\frac{1}{\Delta})}
\end{aligned}
$$

The rate $\sigma_\Delta(s_m)(1+\frac{1}{2}\sigma_\Delta(s_m))$ summarizes the distributional risks for $\Delta > 0$ up to a fourth-order expansion of $\log(1+\sqrt{\sigma_\Delta}(s_m)) + \log(1-\sqrt{\sigma_\Delta}(s_m))$. Exposure to higher moments of the wealth distribution is priced in the incomplete economy, even with myopic investors.

[17] The two aggregation rules produce identical pricing kernels because $[\partial J_{1,j}]^{-1} + [\partial J_{1,-j}]^{-1} = [\partial J_{1,j} + \partial J_{1,-j}]\,(\partial J_{1,j} \partial J_{1,-j})^{-1}$.

Distributional risk becomes negligible as $\Delta W_{1,m}^{-1} \searrow 0$.

Remarks

I. $V_{1,m}$ can be reconstructed $V_{1,m} = \sum_{k\in\{D,U\}} M[\overline{Y}_{m,k}]_{\text{Lucas}}\, Y_{m,k}\, \pi_k$ for each $m \in \{R, G\}$. Because $W_{1,m} = V_{1,m} + \omega_0$, the time-zero prices for payouts at maturity obtain by plugging $V_{1,m}$ into NC, or L.0 for the complete-markets case. The state price kernels $M[\,\cdot\,]$ are a more flexible description of the economy than $V_{1,m}$ in part because, with log utility, the discount rate on a claim to aggregate consumption reduces to the subjective rate of time preference.[18,19]

II. In incomplete markets, investors with a negative shock are compelled to "take the hit," by renormalizing their marginal utility growth rates to be in line with market prices, but at higher individual marginal utility levels. Investors with low private income become poor relative to expectations. Proportionally, the loss of net worth is larger when aggregate productivity is low, holding the level $|\Delta|$ fixed. In the incomplete markets equilibrium, the cross-sectional dispersion of subjective valuations is countercyclical.

2.2.3 Asset Prices with Intermediation

Every investor's time-zero allocations are identical. Ex-post, each type $j = 1, 2$ trades-off the market and the intermediary differently depending on the state of the economy. Consider

[18] In the finite horizon model without intermediate consumption, we set $1 + \beta = 1$.

[19] In an endowment economy with log utility defined over a perishable numeraire c, risk premia on the aggregate claim collapse with a representative agent because $c(s_t)d\log(c(s_t)) = dc(s_t)$ state-by-state. The same would be true in an economy with a durable numeraire and utility defined over wealth, such as ours, if cash was in zero net supply. Notably, in either case, the incomplete markets pricing kernel retains the distributional term when pricing the aggregate claim.

the Euler equations for the low-type $j = 2$, i.e., $n_2 - \Delta_0 = -\Delta$. During a recession $m = R$,

$$\begin{aligned} \partial J_{1,j} \frac{V_R}{V_0} - \mathbb{E}\left[\theta_2(\mathbf{k}^+, k_j, s_{R,k}))^{-1} Y(s_{R,k})\right] &= 0 \\ \partial J_{1,j} V_R - \mathbb{E}\left[\theta_2(\mathbf{k}^+, k_j, s_{R,k}))^{-1} Y(s_{R,k})\right] &\leq 0 \end{aligned}$$

where now $\partial J_{1,j} = [W_{1,j}(\mathbf{k}^+, k_j)]^{-1}$. Consider the high-type $j = 1$ in a recession,

$$\begin{aligned} -\partial J_{1,j} \frac{V_R}{V_0} + \mathbb{E}\left[\theta_1(\mathbf{k}^+, k_j, s_{R,k}))^{-1} Y(s_{R,k})\right] &\geq 0 \\ -\partial J_{1,j} V_R + \mathbb{E}\left[\theta_1(\mathbf{k}^+, k_j, s_{R,k}))^{-1} Y(s_{R,k})\right] &= 0 \end{aligned}$$

whose incentives for trade in each institution are the complement of the low type. The high income investor needs to acquire more risky claims and in a recession they are cheaper on the market. In contrast, the low type must liquidate the claims, and can turn each share into more cash by pulling bank funds. The caveat is that the high type must also not withdraw, but she will never withdraw given her portfolio needs, unless limited liability is jeopardized.

Recall that $\mathbf{k}^+$ is the equilibrating variable for aggregating bank policies. Equating the marginal conditions from the two types gives

$$\frac{\mathbb{E}\left[\theta_1(\mathbf{k}^+, k_1, s_{R,k}))^{-1} Y(s_{R,k})\right]}{\partial J_{1,1}} = V_0 \frac{\mathbb{E}\left[\theta_2(\mathbf{k}^+, k_2, s_{R,k}))^{-1} Y(s_{R,k})\right]}{\partial J_{1,2}} \qquad \text{(Eq.1)}$$

Equation Eq.1 says that market forces relax the imposition on low types that their marginal utility growth equal that of the high type, given their unexpectedly high marginal utility level today. The Euler equation ensures the low type is not raising marginal utility today to make this happen, so the scale factor corresponds to lower marginal utility of wealth tomorrow, i.e., higher ex-post low-type wealth. The relaxation factor is the scale $V_0 < 1$.

The single-period pricing kernel can be written in terms of the Lucas kernel as in the

incomplete markets economy. The kernels are

$$\mathbb{M}[(s_m)]_{\text{IAP}} = [M[(s_m)]_{\text{Lucas}}]\, e^{\sigma_\Delta\left[1-o(\frac{1}{\Delta})\right](s_m,\mathbf{k}^+)}$$

The exponential terms parametrize an equivalent change of measure that distinctly characterizes the asset pricing implications of our financial intermediary.

The Price of Risk

At time-zero, the economy optimally reorganizes to include intermediation. Relative to complete markets, price levels are lower in equilibrium because of a decrease in the over-saving propensity. In addition to the level effect, a covariance effect activates when nontradeable income *levels* are conditionally independent of aggregate output in the time-series.[20] We model nontradeable income additively which, in conjunction with independence from the aggregate state, satisfies this criterion. Other possibly dependent specifications can also be tailored to violate Krueger and Lustig (2010).

To illustrate, consider $n_L < \mathbb{E}[n] < n_H$ and evaluate the excess Euler equations prior to trade. These are equalities at the allocations in expectation. The thought experiment is to consider the effect of news about private income only

$$\begin{aligned} -[\partial J_{L,t}]^{-1}\mathbb{E}[\partial J_{L,t+1}]r_f + [\partial J_{L,t}]^{-1}\mathbb{E}\left[\partial J_{L,t+1}\frac{Y_{t+1}}{P_t}\right] &< 0 \\ -[\partial J_{H,t}]^{-1}\mathbb{E}[\partial J_{H,t+1}]r_f + [\partial J_{H,t}]^{-1}\mathbb{E}\left[\partial J_{H,t+1}\frac{Y_{t+1}}{P_t}\right] &> 0 \end{aligned}$$

With the additive private income specification, convex marginal utility induces not simply rebalancing incentives from changes in wealth, but also relative over-and-under valued types of assets delineated by risk. In contrast, consider a private income process N_j that scales output Y to determine the income level $n_j = N_j Y$, subject to the appropriate goods clearing

[20] This violates the criteria for risk-indifference in Krueger and Lustig (2010).

protocol. The same thought experiment generates the above Euler equations that are still equalized after realizing n. The reason is that with homogeneous preferences, multiplicative income factors out just like aggregate wealth, so that marginal effects from income shocks are constant across asset types.

The Wealth Distribution

We analyse the ex-post wealth distribution in the banking economy against benchmark complete and incomplete markets wealth distributions. Naturally, the complete markets ex-post wealth distribution is degenerate with all mass located at the level of realized output $Y(s_{m,k})$ in each state $s_{m,k}$. A contrasting limit is given by the incomplete markets model, where the wealth distribution is bimodal in each state of aggregate cash flows. The distribution has two atoms of equal mass separated by length 2Δ for each $s_{m,k}$, but the absolute horizontal position of the atoms moves with $Y(s_{m,k})$.

2.2.4 Organizational Implications

We state a handful of implications for industrial organization in the financial sector implied by the benchmark case of our theory.

Corollary 2.2.2 (Banks are always Capitalized) $b_0 > 0$ *and* $\mathbb{P}(\omega(j) = b_0,\ j = 1, 2) < 1$ *Investors allocate strictly positive wealth levels to bank financing, and claim the bank's residual assets with strictly positive probability.*

Corollary 3.1 follows from proposition 2.3. Investors always allocate strictly positive wealth to bank formation initially, $b_0 > 0$. Ex-ante, each investor places positive probability on the event: "retain exposure" to bank assets. Ex-post, the population mass $\pi_1 > 0$ holds bank

capital optimally, in bad times. The bank is always formed ex-ante, and always capitalized ex-post.

Remark This result fails for some changes of model assumptions. If the distribution of private income is not deterministic and $\mathbb{P}(\pi_1 = 0) > 0$, the bank may not be capitalized ex-post. If the distribution of aggregate output incorporates disasters, i.e., realized output of $Y_{\min} = \epsilon > 0$ has $\mathbb{P}(Y_{m,k} = Y_{\min}) > 0$, the bank may not be formed ex-ante.

Corollary 2.2.3 (No Market Segmentation) *Banks coexist with markets: the demand for risky asset intermediation does not rely on restricted access to markets.*

2.2.3 follows from corollary 2.2.2. Investors demand intermediation because risky assets are a necessary input for liquidity production.

Corollary 2.2.4 (Variation in the Investment Opportunity Set) *The demand for financial intermediation is distinct from hedging demand. Myopic investment policies that are indifferent to shocks that impact future productivity nonetheless finance the bank.*

Corollary 3.3 follows immediately from log preferences. Myopic investors prefer one-step ahead contingency plans to mitigate liquidation expenses.

Random Capital Structure The bank has a *random capital structure* in the sense that at any time t assets A_t are financed by a combination of debt and capital that is not known until $t+1$ when population rollover policies are observed and exposures net of liabilities can be computed. The residual claims are risky not only because the cash flows generated by assets are risky but also because the amount of debt financing that survives until residual payments are made is uncertain.

2.3 Conclusion

The formation of a bank improves welfare over an incomplete markets economy when a large risk-averse population faces uninsurable shocks to net worth. Narrow banking precludes this mechanism in its strictest form. Relatively risky assets facilitate the creation of liquidity and hence the impetus for the risk sharing through banks. Bank financing does not in general achieve the first-best allocation because bank liabilities do not allow digital contingencies to be assigned ex-ante. Several interesting implications emerge. We show banks are always capitalized under the assumptions of our economy, even for myopic investors. A corollary to this is that banks and financial markets always coexist in this economy, even when investors have unrestricted access to both. This prediction has evaded a long literature in corporate finance and is worthy of further scrutiny.

The principal thrust of the preceding investigation was to link dynamics of institutional capital structure to the preferences of investors making decisions on the margin who ultimately possess the wealth in the economy. The mechanism can be understood through a key intuitive ingredient that distinguishes a narrow class of institutions including banks from others: the *random capital structure*. Why is bank capital structure risky and why is this unique to banks? Within our stylized economy the answer is: bank capital structure is risky because of the production of liquid liabilities, and only institutions that produce liquidity in the same way exhibit the same patterns in equilibrium. Production of liquid liabilities is a function unique to banks and dealers, each of which contribute uniquely to our understanding of asset prices.

On any day, the assets on the bank's balance sheet are funded by a mix of equity and liabilities. But the mix is uncertain ex-ante. Before any obligatory or residual payments are made, liquidity holders can convert their investment to the numeraire corresponding to the time of their investment. When they do this, the component of the risky assets that their investment effectively financed is transferred pro-rata to the residual owners, who then become liable for the numeraire payment. Based on the investments today, residual claimants face

ex-post leverage restrictions on their cash flow - the same asset has an effective leverage ratio for each state tomorrow in general. Bank and dealer balance sheets are unique in that, in addition to the cash flow risk from the assets, owners bear the risk embedded in the demand options of other liability holders.

The allocations and prices are characterized in several stylized economies for comparison. Asset prices reflect the quality of risk-sharing in the cross-section of the investment population. In each ex-post contingency, the wealth-weighted marginal valuations of investors determine the realized leverage ratio. As a result, changes in bank capitalization rates measure innovations to the average marginal value of wealth in the cross-section of investors. The risk sharing is improved over incomplete markets, which means a reduction in over-savings reflected in the stochastic discount factor.

The implications contribute to understanding the empirical successes of intermediary asset pricing tests. The model provides specific additional implications that can help reject this or other theoretical proposals. A discussion of the implications and evidence are discussed in the appendix. It is useful that the candidate explanation presented in this paper is an outcome in a general equilibrium with minimal primitive assumptions. Indispensable assumptions are few: investors are risk averse and experience uncertainty about the present value of their lifetime productivity, and the markets for the idiosyncratic shocks to valuations are not operational. We assume the income shocks are purely distributional, but the results do not depend on this assumption. The theoretical implications in this paper are transparently prone to rejection by new empirical tests, while exiting models rely on stark assumptions to replicate existing empirical evidence at the expense of generating productive testable implications.

Chapter 3

A False Sense of Security?

An Empirical Investigation of Modern Banking in the United States

with Professor Pryiank Gandhi (Notre Dame University) and Professor Alberto Plazzi (The Swiss Finance Institute)

Extant models of financial intermediation suggest banks diversify out of traditionally *core* activities to improve risk-taking and profitability margins. Diversified banks appear to benefit from "coinsurance." These banks are more profitable, less financially constrained, and supply more credit. However, diversification benefits accrue during periods of normal to high economic growth. Diversified banks are more exposed to systematic risk: their lending is more sensitive to macroeconomic conditions. These banks are also more prone to the risk that nominally correlated activities become highly correlated in bad times. Moreover, we find that banks with higher probability of financial distress diversify more aggressively, suggesting the subset of banks most exposed to systemic shocks are the least equipped to handle them. Our findings are relevant for understanding the optimal scope of banking activities and highlight a novel channel through which noncore banking activities impact credit supply and the real economy.

3.1 Introduction

Traditionally, banks have engaged in just two distinct types of activities: deposit-taking and lending. Modern banks, alternatively, engage in a myriad of activities including trading, brokerage, investment banking, market-making, advisory, underwriting, insurance, and venture capital. Income from these additional activities, once viewed as non-core for banks, has increased significantly, and now accounts for the majority of total income of all U.S. banks.[1]

Conventional wisdom suggests that, for nonfinancial firms, diversification destroys value and adversely affects firm performance Thus, at first sight, U.S. banks' increasing diversification is puzzling. However, banks are different from nonfinancial firms in many respects. Banks are typically highly levered. Banks also serve as delegated monitors for borrowers. Extant literature suggests that, due to these differences, banks should actually strive to be as diversified as possible. Diversification can reduce banks' chance of costly financial distress . It can also make it cheaper for banks to achieve credibility in their role as screeners of borrowers . Finally, since several non-core activities are related to (i.e. require similar "financial" skills as) banks' traditional activities, horizontal diversification into non-core activities may generate a diversification premium for banks.

Empirically, significant controversy in the literature still remains about the effects of banks' diversification into non-core activities on their operations. Regulators and bankers also disagree regarding the benefits of diversification . To our knowledge, existing literature has not examined what effect, if any, does banks' diversification have on their core intermediation capabilities (i.e. lending or credit supply).

In this paper we examine the drivers and consequences of U.S. banks' diversification into

[1]For all U.S. banks, non-core income accounted for only 18% of total income in 1988 but was 55% in 2014.

non-core activities.[2] We address two central questions: (a) "What are the determinants of banks' decision to diversify into non-core activities?" and (b) "What effect does such diversification have on banks' core intermediation capabilities?". We study these questions using a panel data set of about 3,000 unique U.S. banks over the 1987 to 2014 period. The sample includes the 2007 to 2009 financial crisis and therefore allows us to analyze the effect of diversification on bank's performance during all three types of market regimes – pre-crisis, crisis, and post-crisis.

We begin by analyzing the ratio of non-core to total income to document the extent of diversification into non-core activities by all U.S. banks. Over 1987 to 2014, non-core income accounts on average for 37% of total income. This ratio has increased steadily over time from 18.06% (in March 1987) to 54.96% (in March 2014). However, the extent of diversification varies considerably in the cross-section. The ratio of non-core to total income ranges from nearly zero to 80% across individual banks, and is highest for largest banks, defined as those in the top quintile by total book value of assets.

We use the cross-sectional heterogeneity in the degree of diversification at U.S. banks to investigate which factors drive a bank's decision to diversify into non-core activities. That is, using a panel regression framework we relate a bank's extent of diversification into non-core activities to a large set of bank-level and aggregate macroeconomic characteristics. We find that size, as measured by total assets, is significantly positively related to the extent of diversification. Thus, large banks are found to diversify more. The banking literature tends to presume that diversification and size go hand in hand. Our results demonstrate empirically that this presumption is valid. As predicted by theories of delegated monitoring, banks with a high distress costs are also more likely to diversify. Diversification is more pronounced for banks facing higher costs of raising external capital, such as banks with lower deposit ratios and privately-owned banks. Consistent with theories of agency costs of free cash flow, we

[2]Throughout this paper, diversification refers to banks' decision to participate in non-core activities and not the diversification of their loan portfolios.

also document that banks' likelihood to diversify increases in profitability. Finally, we find that managerial incentives, measured by the sensitivity of executive compensation to bank's stock price, are an important determinant of bank diversification. Overall, our findings are consistent with traditional models of financial intermediation and risk management, that suggest banks diversify to lower (idiosyncratic) risk and the cost of delegated monitoring.

Next, we examine whether diversification into non-core activities generates synergistic benefits for banks. For this, we look at the correlation between core and non-core income. For all U.S. banks, over the full sample the correlation between core and non-core income is negative at -0.40, and peaks at -0.72 in the pre-crisis period.[3] Such economically and statistically significant correlations suggest that diversified banks on average benefit from "coinsurance", defined as an imperfect correlation between cash flows generated by different subsidiaries. However, we also observe that the distribution of coinsurance varies across banks and over time as nearly 40% of banks show positive correlation at some point in time. Thus, not all banks in our sample are able to (or choose to) benefit from such coinsurance at all times.

In imperfect markets, coinsurance in bank's activities can have real effects on banks' operations, if it helps banks avoid deadweight financial costs. Many of these costs arise following low internal cash flow realizations (i.e. "financial shortfalls"). For example, consider the deadweight (transaction) costs of raising external finance. If external finance is costly, banks facing financing shortfalls would forego profitable investment opportunities. Coinsurance enables a diversified bank to transfer resources from cash-rich units to cash-poor units in some states to avoid financing shortfalls that standalone banks cannot avoid on their own. Thus, diversified banks should miss (forego) fewer investment opportunities, supply more credit, and be more profitable. These effects should depend on the extent of diversification

[3]We also separately analyze the role of trading income, as it accounts for a large fraction (about 40% in 2014) of non-core income and, in the aftermath of the credit crisis of 2007, was considered as its most controversial component. We find that trading income has a large negative correlation of about -0.77 with core income over the full sample, and -0.86 in the pre-crisis period. We define the times series of core, non-core, and trading income explicitly in our description of the data.

and coinsurance among a diversified banks' income streams.

We test these predictions by exploiting the significant variation in the extent of diversification and coinsurance in the cross-section of U.S. banks. Indeed, better coinsurance is associated with lower financial constraints and higher average credit supply. In particular, banks with a higher proportion of non-core to total income and better coinsurance (i.e. a more negative correlation between income streams) are able to issue more credit after controlling for capital, deposit growth, profitability and growth opportunities. The effect is statistically as well as economically significant, as a 1% increase in coinsurance is accompanied with a 2.7% increase in credit supply. Further, consistent with the idea that diversification and coinsurance help relax financial constraints, we find that diversified banks with better coinsurance are more profitable, issue more dividends, and that their credit supply is less sensitive to changes in internal cash flows.

Taken together, these findings imply that diversification is beneficial. Banks that should theoretically benefit from engaging in non-core activities are the ones that tend to diversify and benefit the most from it. This conclusion, however, begs the obvious question why don't all banks participate (or participate more) in non-core activities. A natural explanation is that the additional investments required to prepare and support diversification (such as hiring skilled personnel, investing in systems that monitor and report risk across subsidiaries, etc.) are costs that some banks find prohibitive, especially in times when profit margins on non-core activities are low.

In this paper, we provide evidence of a complementary, risk-based channel that may outweigh the benefits of bank diversification and suggests that diversified banks may be operating under a *false sense of security*. For the subset of traded banks, we find that diversification reduces idiosyncratic risk, but increases systematic as well as total risk. In particular, diversification increases banks' exposure to "correlation" risk. Correlation risk arises because of unexpected changes in the relationship between core and non-core incomes, which tend to coincide with times of aggregate shocks. In other words, even if banks design

their non-core activities to provide diversification benefits, they expose themselves to unexpected shocks in correlation that may override the hedging role of this additional income. This is exactly what occurred during the crisis period of 2007 to 2009 when the correlation between core and non-core income turns positive (and it again later becomes negative in the post-crisis period). Thus, diversification benefits are limited to good times and the diversification and coinsurance mechanisms (whereby banks rely on incomes generated by diversified business units) can and does break down, precisely when it's most needed.

To the extent that exposure to correlation (systematic) risk is priced in the market, diversification raises the cost of capital of diversified banks and this tends to happen in times when the economy faces aggregate shocks. Consistent with this argument, we show that the adverse effect of diversification on banks' cost of capital in bad times impacts their credit supply. Moreover, the sensitivity of credit supply to macroeconomic conditions (measured by GDP growth) is highest for banks that diversify the most, and are therefore more exposed to changes in the investment opportunity set (i.e., they are more systematically risky).

Are bank managers aware of these costs of diversification? If yes, why do they still choose to diversify bank activities? One possibility is that diversification allows banks to grow rapidly beyond a certain size threshold. Large banks benefit from a lower cost of capital due to implicit and explicit government guarantees. Hence, it is possible that managers diversify to receive benefits of implicit guarantee and this exceeds the costs of diversification outlined above. In fact, we show that benefits of implicit government guarantees (as measured by the loadings on the size factor of Gandhi and Lustig (2015) increase monotonically in diversification.

Our paper relates to the growing literature that analyzes how diversification into non-core activities affects banks' operations. Our paper differs from these other papers as it is the first to use information not only on the proportion, but also on the correlation between core and non-core income. Thus, we present evidence for why and under what circumstances does diversification affect bank operations. We are also the first to relate diversification to

idiosyncratic and systemic components of bank risk and show that U.S. banks' diversification increases exposure to correlation risk.

Our findings are quite timely given the regulatory debate on optimal scope of bank activities. We highlight novel channels through which diversification impacts banks' credit supply and therefore the real economy. In the aftermath of the credit crisis, existing regulatory proposals in the U.S. and many other countries advocate the separation of deposit-taking and lending from many activities that are traditionally considered as non-core for the banking sector ("ring-fencing"). These proposals are opposed on the grounds that diversification into non-core activities helps banks manage cash flow risk and improves their overall safety. Our paper provides additional information that academics, regulators and practitioners can use to assess the costs (and benefits) of banks' participation in non-core activities.

3.2 Data description

3.2.1 Sample selection

We collect balance sheet data from the 'Report for Condition and Income' (henceforth the Call Report) required to be filed by all FDIC-insured bank holding companies (henceforth banks). In the U.S., banks with total book value above $500 million file this report quarterly whereas other banks file this report semi-annually. We restrict our sample to banks which file the Call Reports quarterly. This restriction implies that our sample includes 2,978 unique banks (i.e. 67%) of 4,460 banks in existence in the U.S. as of December 2014. The benefit is that it allows us to analyze financial data at the highest frequency possible. Our sample includes large and medium banks which collectively account for more than 90% of book value of assets for all U.S. banks. Since diversification into non-core income is positively correlated with size, as our subsequent analysis shows, we expect our sample to be well representative of the average bank that engages in such activities. Call Report data starts in September 1987, and this determines the start date of our sample.

A typical bank in the sample owns multiple subsidiaries that provide commercial banking or other financial services. Banks can also have stakes in non-financial firms although such ownership cannot exceed 5% of the non-financial firm's outstanding equity. For Call Reports, a bank is required to aggregate data only for subsidiaries that provide commercial banking or other financial services. Thus, by definition our data excludes non-financial subsidiaries owned by a bank, if any.

A drawback of our aggregated data is that we are unable to say how diversification within an individual commercial banking subsidiary impacts its operations. However, since most banks with several subsidiaries manage capital centrally our aggregated data provides the ideal empirical setting for our analysis.[4]

3.2.2 Measuring the extent of bank diversification

We begin by computing the income from core, non-core, and trading activities. Income from a bank's core activities is simply the total interest income, adjusted for loan-loss provisions, less any interest income generated by a bank's trading assets. We adjust for loan-loss provisions as these are an estimate of expected losses on a bank's loan portfolio. Hence, our definition of core income results in a measure of actual cash flows that a bank expects to receive from its core activities.[5]

Income from non-core activities is the total non-interest income generated by the bank. We also separately define the time-series of the most controversial type of non-core income i.e. trading income. The Call Reports lists several items that could serve as a proxy for a bank's trading income. The obvious starting point is the banks reported trading revenues. To this, we add interest income from trading assets. We also add realized gains or losses

[4]In addition, for all banks in our sample, traded equity prices reference the entire firm, and not individual subsidiaries. Were we to use data only for individual commercial banking subsidiaries, we would be unable to carry the analysis that relies on traded equity.

[5]Banks may manipulate provisions to smooth earnings. Our central empirical results are robust to computation of core income with or without adjustment for loan loss provisions.

from held-to-maturity and available-for-sale securities. Note that, unless explicitly specified, income from non-core activities will always include trading income.[6]

To summarize, our measures of core, non-core, and trading incomes in a given bank-quarter are given by:

$$\begin{aligned} \texttt{Core Income} &= \text{Interest - Interest on trading assets - Provisions} \\ \texttt{Non-Core Income} &= \text{Non-interest (including Trading)} \\ \texttt{Trading Income} &= \text{Trading revenue} + \text{Interest on trading assets} + \text{Realized gains/losses} \end{aligned}$$

Armed with these series, we compute various measures of the extent of banks' diversification into non-core activities. First, we look at the ratio of non-core income and separately, of its trading income component to total income. We denote these ratios for bank j in quarter t by $NonCore_{j,t}$ and $TRA_{j,t}$, respectively. The ratios reflect how much of a bank's overall income originates from non-core activities. Banks with low ratios possess a lower degree of diversification.

These income ratios, however, provide only a partial view of the potential benefits from diversification, as they ignore how the two income streams co-move. Portfolio theory implies that lower correlations allows one to achieve better risk-return combinations (holding volatilities fixed). For this reason, we next compute the sample correlation between core and non-core income, and treat it as an inverse measure of the extent of diversification synergies (i.e. coinsurance) provided by non-core activities. Higher correlation is associated with lower diversification synergies and coinsurance benefits. In other words, banks with better coinsurance are those with lower (or negative) values of correlation. We denote the unconditional correlation between core and non-core income for bank j by $\boldsymbol{\rho_j}$.

[6]Due to changing reporting requirements, some of the Call Report items used for computation of core, non-core, and trading incomes are not comparable across quarters. We follow instructions provided by the Chicago Federal Reserve Bank to form consistent time-series for all our variables, available at http://www.chicagofed.org/webpages/banking/financial_institution_reports/bhc_data.cfm.

We also construct a conditional version of correlation, which we denote by $\boldsymbol{\rho_{j,t}}$. At each quarter t, $\boldsymbol{\rho_{j,t}}$ is computed in a rolling fashion using income data over each of the 20 quarters preceding and including t, i.e. from $t-20$ to t. This conditional measure allow us to track time variation in a bank's degree of coinsurance, and to relate it to lending decisions and profitability. For consistency and direct comparability, in the empirical tests that use $\boldsymbol{\rho_{j,t}}$ we also average the proportions ($NonCore_{j,t}$ and $TRA_{j,t}$) over the same 20-quarter period.

3.2.3 Descriptive statistics

Table 3.1 presents summary statistics for core, non-core, and trading incomes expressed as a percentage of total income. The last row in each panel reports the fraction of trading assets to total book value of assets. Panel A reports statistics for the aggregate U.S. bank sector as well as banks sorted into five groups by size. To obtain the time series for the aggregate bank sector, we sum up the incomes across all U.S. banks in a given quarter. Similarly, to get time series for the size-sorted groups of banks, in each quarter, we first rank banks into quintiles based on the total book value of assets. Group 1 refers to banks in the lowest total book value of assets quintile. We then sum up the incomes across all U.S. banks in a given quarter in a given group. We report the (time-series) mean and standard deviation of these series.

Over the sample period, for all banks in the U.S., we see that non-core income accounts for nearly one-third of total income. Further, trading income accounts for 13% of total income and trading assets are on average about 7% of total U.S. bank assets. There is, however, considerable time-series variation, and these ratios have been generally increasing over the second half of the sample period. In levels, quarterly core income has grown from \$35 billion to \$79 billion over 1987 to 2014. In contrast, non-core income has grown from \$7 billion to \$56 billion over the same period. We also note that the ratio of non-core and trading incomes to total incomes are positively related to size. Banks in the lowest size quintile generate on average about 18% of their income from non-core and about 5% of their

Table 3.1: Summary statistics for the proportion of core, non-core, and trading income

Notes: This table presents summary statistics for core, non-core, and trading incomes (all normalized by total income) as well as trading assets (normalized by total book value of assets). Panel A reports the statistics for the aggregate quarterly series, and for the size-sorted groups of banks. In each quarter, we sort banks into 5 groups based on total book value of assets. Group 1 refers to banks with the smallest total book value of assets. Panel B reports statistics for the cross-sectional distribution of the ratios across all banks in our sample. Quarterly data, September 1987 to December 2014.

Panel A: Aggregate

	All banks		Size-sorted banks				
	Mean	**Std**	1	2	3	4	5
Core	63.00	15.31	82.43	84.26	82.57	79.91	61.92
Non-core	37.00	15.31	17.57	15.74	17.43	20.09	38.08
Trading	12.96	10.87	5.35	5.40	5.74	5.47	13.33
Trading assets	7.12	3.90	0.03	0.04	0.05	0.10	7.56

Panel B: Cross-section

Core	84.18	86.91	10.94	22.74	79.97	91.51	101.02
Non-core	15.82	13.09	10.94	−1.02	8.49	20.03	77.26
Trading	6.01	0.67	9.76	−3.64	0.00	8.85	60.66
Trading assets	0.08	0.00	0.40	0.00	0.00	0.00	5.14

income from trading activities. These figures more than double for the largest banks, with a ratio of non-core (trading) to total income of about 38% (13%).

Panel B of the table presents the cross-sectional distribution of the diversification ratios. While for the average bank in the sample non-core activities account for about 16% of overall income, the dispersion is quite large with a standard deviation of nearly 11%. In addition, the interquartile range varies from 8% to 20%. Over our sample period, the ratio of non-core to total income peaks at 77%. Similar conclusions apply to the ratio of trading to total income. While the average is 6%, the distribution is highly skewed. A quarter of the banks exhibit no income in a given quarter, while for some others the ratio peaks at 60%.

To understand how the cross-sectional distribution evolves over time, we plot the proportion of banks for whom non-core income accounts for less than 10% of total income in any given quarter. As is clear from the graph, this proportion has steadily decreased over time from above 50% in the early part of the sample to as low as 13% in the most recent period. The Gramm-Leach-Bliley Act – which was passed by the U.S. Congress in 1999 and allowed bank holding companies to participate in insurance, securities, and investment banking activities – is accompanied by a decrease by about 30% in this proportion. This evidence confirms our statement that U.S. banks have increasingly diversified into activities

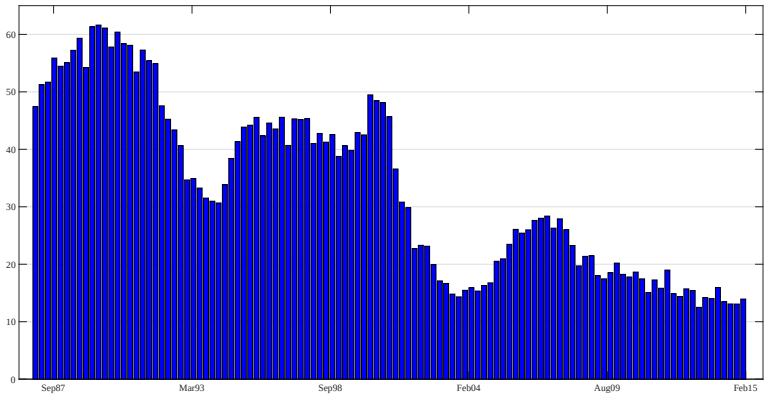


Figure 3.1: Banks with low non-core income

This figure plots the proportion (in %) of banks for whom non-core income is less than 10% of total income. Quarterly data, September 1987 to December 2014.

historically considered non-core for the banking sector.

Table 3.2 presents the summary statistics for the correlation between core, non-core, and trading incomes. Panel A presents the correlation between core, non-core, and trading incomes for the aggregate U.S. bank sector across different sample periods. In the first column correlations are computed over 1987 to 2014. For this period, the correlation between core and non-core income is negative at -0.40, while that between core and trading income is even more pronounced at -0.77. Thus, overall it appears that banks' diversification into non-core activities provides significant synergistic benefits. However, such diversification synergies vary considerably over time. The next three columns in Panel A break down the correlation between the pre-crisis period (1987:Q3-2006:Q4), crisis period (2007:Q1-2009:Q1), and post-crisis period (2009:Q2-2014:Q4). While the correlation between core and non-core incomes is negative in the pre- and post-crisis periods at -0.72 and -0.31 respectively, it turns positive at 0.34 (albeit not statistically significant) during the recent financial crisis. The impact of the crisis is even more pronounced on the correlation between core and trading incomes. Many of the synergies even in banks' core businesses broke down during the recent crisis. Overall, the evidence that emerges from these statistics lends preliminary support to the following mechanism: U.S. banks benefit from diversification synergies as they increase

reliance on non-core activities. However, any coinsurance provided by non-core activities can break down, and this can happen precisely when such coinsurance is needed most.

Panel B reports summary statistics for the distribution of the time-varying, conditional correlations $\boldsymbol{\rho_{j,t}}$ (defined above). A unit of observation in this panel is a bank-quarter combination. The average correlation between core and non-core income is negative at -0.05 and that between core and trading income is even more pronounced at -0.20. However, there is significant variation in the extent of coinsurance between core and non-core income in the cross-section of U.S. banks, as the standard deviation is 0.40. For about 55% of the observations (i.e. bank-quarters) the correlation between core and non-core (or trading) incomes is negative. While no bank has zero non-core income (and hence a zero correlation between core and non-core income), for about 18% of the observations in our sample we observe zero trading income (and hence zero correlation). We exploit this fact, as well as the fact that some banks begin participating in trading activities, in our robustness tests below to provide further causal evidence on the impact of diversification on banks' core intermediation capabilities.

The picture that emerges from these statistics is that: i.) non-core activities account for about one-third of total income during the whole sample; ii.) there is considerable variation over time and in the cross-section in the extent of diversification and coinsurance; iii.) non-core income is negatively related to core income, although this correlation turns positive during the crisis.

The next section discusses how we use this wealth of information to investigate what determines the extent of bank's diversification into non-core activities, and what effect does this have on bank's operations.

Table 3.2: Summary statistics for correlation of core, non-core and trading income.

Notes: This table presents summary statistics for the correlation between core and non-core or trading incomes, where all incomes are normalized by total income. Panel A reports the correlation for the aggregate quarterly series over different sample periods: full sample; pre-crisis period (1987:Q3-2006:Q4); crisis period (2007:Q1-2009:Q1); and post-crisis period (2009:Q2-2014:Q4). Panel B reports statistics for the cross-sectional distribution of the conditional correlation $\boldsymbol{\rho_{j,t}}$. 'Negative' is the fraction of negative correlations; 'Zero' is fraction of zero correlations (i.e. no non-core or trading income over the 5-year period); and 'Positive' is the fraction of positive correlations. Statistical significance at the 10%, 5% and 1% levels is denoted by *, **, and *** respectively. Quarterly data, September 1987 to December 2014.

Panel A: Aggregate

	Full sample	Pre-crisis	Crisis	Post-crisis
Non-core income	-0.3967^{***}	-0.7235^{***}	0.3441	-0.3081^{***}
Trading income	-0.7721^{***}	-0.8647^{***}	0.3166	-0.5445^{***}

Panel B: Cross-section of $\boldsymbol{\rho_{j,t}}$

	Mean	Std	25th	Median	75th	Neg.	Zero	Pos.
Non-core	−0.0530	0.4064	−0.3735	−0.0628	0.2541	0.5525	0.0001	0.4474
Trading	−0.1996	0.3998	−0.5336	−0.1057	0.0002	0.5670	0.1829	0.2502

3.3 Hypotheses and variables construction

3.3.1 Determinants of banks' decision to diversify

Diamond (1984) develops a theory of financial intermediation in which banks have a comparative advantage over individual savers in their ability to monitor the credit quality of borrowers. An implication of this family of models is that diversification within a bank lowers the cost of monitoring. This is one of several channels through which, unlike nonfinancial firms, banks may find it optimal to diversify. Moreover, this incentive should increase with size.

Pursuing this line of reasoning: if some credit risk is only partially diversifiable, banks may choose not only to hold diversified loan portfolios, but also to diversify into financial activities traditionally considered as non-core. Thus, economies of scale and scope provide incentives for banks to diversify. These arguments suggest that the extent of diversification is driven by bank size, which we measure by the log total value of assets ($Size$).

Generally, the literature on risk management shows that hedging increases firm value by reducing the volatility of firm's cash flows , lowering the expected cost of bankruptcy, expanding the debt-capacity for firms , reducing investment distortions, and helping firms avoid deadweight transaction costs due to costly external finance . Thus, we expect banks facing a higher cost of financial distress or higher cost of external finance to be more likely to diversify their cash flows. We capture the cost of financial distress using the ratio of income variability to expected equity capital, and denote it $Distress\,Costs$.[7] We proxy for the cost of raising external capital by the bank's listing status. Publicly-listed banks can easily tap equity markets to raise capital and may have a lower cost of raising external finance than privately-held banks. Thus, our analysis includes a dummy variable *Listed* that equals one for public banks, and zero otherwise.

It is likely that banks that rely more on deposit financing have access to a (more) stable source of funding. These banks would also face a lower cost of raising external finance (due to deposit insurance), and may therefore have a lower incentive to diversify into non-core activities. To capture this effect, we include a bank's deposit ratio, *Deposit Ratio*, computed as the ratio of deposits to total assets.

Theoretically, high-growth banks are more likely to face financing shortfalls and therefore have a stronger motivation to diversify in order to avoid the cost of raising external capital. For this reason, we consider the average quarterly growth rate of a bank, $Loan\,Growth$, computed as the percentage growth in loans over the past four quarters.

Bank profitability should also impact the management decision to diversify. Managers at more profitable banks will not only have the resources to expand beyond core banking activities, but due to agency costs will also have a propensity to invest to grow banks beyond their optimal size and scope. We measure profitability by income growth, $Income\,Growth$, defined as the quarterly change in net income scaled by total assets.

[7]To be precise, the measure is computed as the variance of quarterly net income to the square of equity capital plus average quarterly net income.

Finally, we control for managerial incentives and agency costs that are embedded in manager's compensation. Theory predicts that the incentive to manage risks (i.e. diversify) should be positively related to the sensitivity of managerial wealth to stock prices, and negatively related to its sensitivity to stock return volatility. For the subset of publicly-listed banks that are covered by the COMPUSTAT executive compensation database, we compute the $Delta$ and $Vega$ of the top management share and option holdings following the procedure in CM (2016). We expect diversification to relate positively to $Delta$ but negatively to $Vega$.

Note that our paper takes the degree of coinsurance (i.e. correlation) between core and non-core income as given, and does not investigate why this varies across banks. The reason for this is data limitations. We do not observe the exact nature of transactions that comprise a particular bank's core, non-core, and trading income. Hence, we are unable to comment on what drives the coinsurance and why this coinsurance varies across banks.[8]

We speculate that the degree of coinsurance between core and non-core income varies across banks and over time as it is a choice (design) variable for bank managers.Further, we assume that bank managers make this choice rationally based on all available information. For example, consider the decision by a bank to participate in proprietary trading activities. Proprietary trading in options and derivatives may allow bank managers to directly control correlation exposures and as a result the degree of coinsurance between core and non-core incomes. In fact, this view would be consistent with Atkeson, Eisfeldt and Weill (2015), who develop a model for a bank that trades derivative securities to hedge heterogeneous exposures to loan default risk. In their model, a bank actively hedges loan default risk via trading positions.

The cross-sectional variation in the degree of coinsurance between core and non-core income could also be driven by costs of banks' participation in non-core activities. These costs include but are not limited to hiring skilled personnel, investing in systems that monitor

[8]Data that allows researchers to comment on this are hard to get even for regulators.

and report risk across subsidiaries, etc, and are difficult to observe for all banks in our sample.

3.3.2 Impact of diversification on banks' operations

We relate our two measures of diversification i.e. the ratio of non-core to total income and coinsurance to five aspects of banks' operations namely, their profitability, financial constraints, credit supply, risk profile, and response to changes in business conditions.

Diamond (1984) argues that diversified banks incur a lower cost of delegated monitoring and hence should be more profitable. The model of financial intermediation in BHW (1995) also suggests that diversified banks generate higher profits for the same level of risk. Further, if bank managers are rational profit-maximizers, non-core activities pursued by banks should, on average, enhance profits. Finally, the banking literature assumes that non-core activities, which are often fee-based, do not expose banks to interest rate risk, improve profitability and reduce earnings volatility. Therefore, we analyze how diversification impacts the level and volatility of bank profitability. Profitability is measured by the ratio of net income to total book value of assets. We denote the level of profitability by P and the volatility of profitability by σP, respectively.

Minton and Schrand (MS) (1999) show that for nonfinancial firms, higher cash flow volatility is associated with lower investment. This is because firms with higher cash flow volatility are more likely to face periods of internal cash flow shortfalls. Evidence in MS 1999 indicates that firms react to these shortfalls not by simply changing the timing of investments to match cash flow realizations but by forgoing investments altogether. If diversification improves profitability and lowers earnings volatility for banks, then diversified banks should provide more credit. Further, in the cross-section this effect should depend on the extent of diversification/coinsurance. In our analysis, we use four separate proxies for credit supplied by banks namely, total credit ($Total$), commercial and industrial credit ($Commercial$), real estate credit ($Real$), and credit provided via commitments ($Commit$). For each bank, all four measures of credit are normalized by beginning of period book value of assets.

Current research often stresses the efficiency-enhancing role of internal capital markets in diversified firms. These studies suggest that internal capital allocations in diversified banks can provide an intertemporal insurance function against financing shortfalls and financial constraints. We investigate the link between diversification and financial constraints faced by banks. We measure the financial constraints status of a given bank by the dividend payout ratio and by the sensitivity of its credit supply (i.e. its allocation of funds or investment) to its internal cash flows. Dividend payout ratio, denoted by DIV, is computed as the ratio of total dividends to net income.

Next, we examine the link between diversification and measures of bank risk. A fundamental implication of modern portfolio theory is that diversification reduces the return variance of a portfolio of financial assets. Applied to banking, portfolio theory suggests that diversification can potentially reduce the probability of failure and make banks less risky. However, due to managerial incentives (especially agency costs) banks can also misuse their diversification advantages and operate with more leverage or pursue riskier activities. For example, if diversification improves profitability, managers at better diversified banks will have the resources to invest to grow banks beyond their optimal size and scope. For our analysis, we use various market based measures of risk derived from equity returns such as total risk, idiosyncratic risk, and systematic risk.

Finally, diversification should also impact how bank credit supply reacts to changes in business (macroeconomic) conditions. Kayshap and Stein (2000) show that the credit supply of banks facing tougher financial constraints is more responsive to changes in business conditions and monetary policy.In addition, the theoretical model of Diamond (1984) also suggests that credit supply of better diversified banks is less sensitive to aggregate shocks. These studies suggest that by relaxing financial constraints and improving liquidity, diversification can provide a direct advantage to banks, and can insulate its credit supply from sudden changes in business conditions. Alternatively, (as suggested above) if coinsurance benefits vary over time and reverse exactly when they are needed most, diversification could

increase bank's cost of capital in times of aggregate shocks, and amplify the sensitivity of bank credit supply to changes in business conditions. To proxy for macroeconomic conditions we use the year-on-year growth of the gross domestic product (GDP). This variable allows us to document how diversification impacts the relationship between bank credit supply and macroeconomic conditions during different market/macroeconomic regimes.

3.4 Empirical results

In this section, we present our main empirical results. We first investigate the determinants of bank's decision to diversify. We then look at the impact of diversification on bank operations.

3.4.1 The determinants of banks' decision to diversify

We relate the extent of bank's diversification in non-core activities to the variables outlined previously using a panel regression. To recap, our analysis includes the following bank-level determinants along with their expected signs in parentheses: $Size$ (+); $Distress\,Costs$ (+); $Listed$ (-); $Deposit\,Ratio$ (-); $Loan\,Growth$ (+); $Income\,Growth$ (+); $Delta$ (+); and $Vega$ (-). In addition, in each regression, we control for four macroeconomic variables that capture aggregate conditions, as these may affect a bank's core/non-core income and hence its incentive to diversify. The level of interest rates ($Level$, measured by the yield on the three-month U.S. T-bill) captures macroeconomic conditions and monetary policy stance. Diversification to avoid a financing shortfall may be particularly relevant when monetary policy is tight or macroeconomic conditions worsen. Positive shocks to interest rates volatility (Vol, measured by the first difference in the volatility of daily changes in the three-month U.S. T-bill in a given quarter) may increase the volatility of a banks' core income and thus induce higher diversification incentives. We also include the term spread ($TSPR$, the difference between the ten-year and the one-year U.S. bond yields) and the credit spread ($DSPR$, the difference between yields on corporate bonds issued by BAA- and AAA-rated

firms in the U.S.) as these are often regarded as proxies for business cycle fluctuations[9].

Table 3.3 presents the estimates for the panel regression as the cross-sectional determinants are progressively included. We observe that bank size is statistically significant across all specifications with the expected positive sign. Size is also economically significant, as a 1% increase in total assets is accompanied by an increase of 3% in the proportion of non-core activities.

Expected distress costs also enter the regression with an expected positive loading. The coefficient is statistically significant at the 1% level for all but the last two specifications. This may be on account of the fact that, in the last two columns, we restrict our sample to banks present in the COMPUSTAT executive compensation dataset. This reduces the number of observations by approximately 90%.

The coefficient on *Listed* is negative at -0.02, and is again highly significant. This estimate implies that, ceteris paribus, the proportion of non-core activities of privately-owned banks is on average 2% higher than that of otherwise identical public banks even after controlling for size. This result is consistent with existing theories of risk management that argue that the propensity to diversify is negatively related with the costs of raising external finance. Similarly, banks with a higher deposit ratio are less likely to diversify into non-core activities, and the effect is strongly significant in specifications 4 to 6. It seems that insured deposits indeed provide a stable low cost source of external funds so these banks are less in the need of diversifying their cash flows.

Turning our attention to loan growth, we see that the coefficient has an unexpected negative sign which turns insignificant in the sample restricted to publicly-listed firms (i.e. columns 7 and 8). High loan growth banks are more likely to face financing shortfalls and therefore have a stronger motivation to diversify to avoid raising external capital. Therefore, the negative coefficient on loan growth in Table 3.3 appears puzzling. A potential explanation

[9]An alternative approach is to control away aggregate conditions altogether using time fixed effects. We verify our results are robust to this alternative.

Table 3.3: Determinants of bank diversification into non-core activities

Notes: This table presents the results of the panel regression of bank characteristics on the bank's decision to diversify into non-core activities. The dependent variable is the proportion of a bank non-core income. Each column presents the results for a separate specification. $Size$ is the log total book value of assets; $Distress\,Costs$ is a measure of expected distress costs; $Listed$ is a dummy variable that equals one if the bank is publicly listed; $Deposit\,Ratio$ is the ratio of deposits to total assets; $Loan\,Growth$ is total loan growth; $Income\,Growth$ is the change in net income by total assets; $Delta$ and $Vega$ measure executive incentives; $Level$ is the 3-month U.S. T-bill yield; $TSPR$ is the term spread; $DSPR$ is the default spread; and Vol is the change in 3-month yield volatility. Quarterly data, September 1987 to December 2014.

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Size$	0.0325^{***}	0.0334^{***}	0.0357^{***}	0.0310^{***}	0.0310^{***}	0.0312^{***}	0.0324^{***}	0.0282^{***}
	(20.44)	(19.69)	(18.59)	(13.01)	(13.02)	(13.14)	(5.31)	(4.69)
$Distress\,Costs$		0.0111^{***}	0.0100^{***}	0.0108^{***}	0.0104^{***}	0.0140^{***}	0.0402	0.0236
		(3.41)	(3.06)	(3.32)	(3.17)	(3.74)	(1.12)	(0.69)
$Listed$			-0.0150^{***}	-0.0159^{***}	-0.0158^{***}	-0.0154^{***}		
			(-3.18)	(-3.43)	(-3.41)	(-3.33)		
$Deposit\,Ratio$				-0.1414^{***}	-0.1413^{***}	-0.1413^{***}	-0.0911	-0.0682
				(-3.35)	(-3.35)	(-3.36)	(-0.83)	(-0.61)
$Loan\,Growth$					-0.0078	-0.0121	-0.0652^{**}	-0.0649^{**}
					(-0.72)	(-1.11)	(-2.25)	(-2.25)
$Income\,Growth$						2.2406^{**}	0.3324	0.0168
						(2.04)	(0.11)	(0.01)
$Delta$							0.0148^{***}	0.0099^{*}
							(2.71)	(1.67)
$Vega$								0.0171^{***}
								(2.38)
$Level$	-0.0334^{***}	-0.0320^{***}	-0.0309^{***}	-0.0309^{***}	-0.0308^{***}	-0.0312^{***}	-0.0192^{***}	-0.0182^{***}
	(-30.05)	(-26.51)	(-24.94)	(-25.40)	(-25.12)	(-24.72)	(-4.41)	(-4.13)
$TSPR$	-0.0280^{***}	-0.0260^{***}	-0.0246^{***}	-0.0238^{***}	-0.0238^{***}	-0.0241^{***}	-0.0041	-0.0033
	(-22.03)	(-17.90)	(-16.60)	(-16.66)	(-16.68)	(-16.64)	(-0.78)	(-0.62)
$DSPR$	-0.0049^{***}	-0.0034^{*}	-0.0030	-0.0060^{***}	-0.0060^{***}	-0.0049^{**}	0.0029	0.0019
	(-2.79)	(-1.75)	(-1.55)	(-2.69)	(-2.67)	(-2.15)	(0.41)	(0.28)
Vol	0.0222^{***}	0.0374^{***}	0.0370^{***}	0.0259^{***}	0.0256^{***}	0.0232^{**}	0.0257	0.0191
	(2.48)	(3.88)	(3.84)	(2.51)	(2.48)	(2.25)	(0.68)	(0.50)
N	$59,941)$	$51,024$	$51,024$	$51,024$	$51,024$	$51,024$	$4,410$	$4,410$
R^2	0.35	0.34	0.34	0.35	0.35	0.35	0.31	0.32

is that in a bank with high loan growth the loan-granting subsidiary may have stronger (political) influence over investment decisions.

Column 6 shows that, as expected, income growth is positively related to the decision to diversify, and that the effect is statistically significant at the 5% level, while it is not statistically significant in the last two columns. Thus, it appears that profitable banks are more likely to use their resources to diversify into non-core activities.

Finally, columns 7 and 8 control for managerial incentives for the subset of publicly-listed banks for whom data is available in COMPUSTAT. We find that the propensity to diversify into non-core activities is positively correlated with the 'Delta' of managerial share and option holdings. The effect is statistically significant at the 10% level or better. However, we find that the propensity to diversify also increases in 'Vega'. This may be on account of the fact that, for publicly-listed banks, the correlation between 'Delta' and 'Vega' is very high at 0.77, which makes it hard to disentangle their effect on banks' diversification decision. We also observe that the extent of diversification varies over time in a predictable manner. It tends to be inversely related to the level of interest rates, the term spread, and the default spread, and positively related to interest rate volatility.

Overall, our results lend support to the extant models of financial intermediation and risk management. Economies of scale and scope (i.e. size) play a very important role in banks' decision to diversify into non-core activities. The incentive to manage risks via diversification is stronger for banks facing a high cost of raising external finance and those with higher distress costs. Further, as predicted by models of agency cost, profitable banks are more likely to invest to grow banks beyond their optimal size and scope, and the propensity to diversify is also positively correlated with the 'Delta' of managerial shares and options holding.

3.4.2 Impact of diversification on banks' operations

For analyzing the impact of diversification on banks' operations, we note that since data on instantaneous correlation between core and non-core income is not available, in a given quarter, for a given bank, we measure coinsurance using data over the past 5 years (i.e. 20 quarters). For consistency, all other (i.e. dependent and control) variables in the regressions in this section are also measured over the same 5-year period.

Impact of diversification on bank profitability

We test the relationship between bank diversification/coinsurance and bank profitability by estimating the following pair of panel regressions:

$$\begin{aligned} P_{j,t} &= \text{const.} + TFE + \beta_1^{PR} NonCore_{j,t} + \beta_1^{\rho} \boldsymbol{\rho}_{j,t} + \beta_1^{DG} DG_{j,t} + \beta_1^{CAP} CAP_{j,t} + \epsilon_{j,t} \\ \sigma P_{j,t} &= \text{const.} + TFE + \beta_2^{PR} NonCore_{j,t} + \beta_2^{\rho} \boldsymbol{\rho}_{j,t} + \beta_2^{DG} DG_{j,t} + \beta_2^{CAP} CAP_{j,t} + \varepsilon_{j,t} \end{aligned} \quad (3.1)$$

Here, $P_{j,t}$ denotes total income scaled by total assets for bank j in quarter t, averaged over the 5-year period over which we measure $\boldsymbol{\rho}_{j,t}$ and $NonCore_{j,t}$. The standard deviation of cash flows, $\sigma P_{j,t}$, is also computed over the same period. We expect that higher diversification and coinsurance during a given 5-year period is associated with higher overall cash flows and lower volatility of cash flows. In the regressions, we control for the average quarterly growth in deposits (DG), and the average ratio of book value of equity to book value of assets (CAP), computed over the same 5-year period. The regressions also include time fixed effects. Standard errors are clustered by bank.

Estimates of the models are given. The table shows that banks with a higher proportion of non-core income are not only more profitable, but their cash flows are also more volatile. Also note that the impact of diversification (i.e. $NonCore$) on average profitability is not significant once we control for capital and deposit growth rates.

In contrast, higher coinsurance is associated with higher overall cash flow levels and lower

Table 3.4: Impact of diversification on bank profitability

Notes: This table presents results of the following panel regressions:

$$\begin{aligned} P_{j,t} &= \text{const.} + TFE + \beta_1^{PR} NonCore_{j,t} + \beta_1^{\rho} \boldsymbol{\rho}_{j,t} + \beta_1^{DG} DG_{j,t} + \beta_1^{CAP} CAP_{j,t} + \epsilon_{j,t} \\ \sigma P_{j,t} &= \text{const.} + TFE + \beta_2^{PR} NonCore_{j,t} + \beta_2^{\rho} \boldsymbol{\rho}_{j,t} + \beta_2^{DG} DG_{j,t} + \beta_2^{CAP} CAP_{j,t} + \varepsilon_{j,t} \end{aligned}$$

Here, $P_{j,t}$ denotes total income scaled by total assets for bank j in quarter t, averaged over the 5-year period over which we measure $\boldsymbol{\rho}_{j,t}$ and $NonCore_{j,t}$. The standard deviation of cash flow, $\sigma P_{j,t}$, is also computed over the same period. DG, and CAP are respectively the quarterly growth rate of deposits, and the ratio of book value of equity to total book value of assets, averaged over the last 5 years. TFE are time fixed effects. Statistical significance at the 10%, 5% and 1% levels is denoted by *, **, and *** respectively. In parentheses we report t-statistics based on standard errors clustered by bank. Quarterly data, September 1987 to December 2014.

	P	σP	P		σP
$NonCore$	0.0009***	0.0009***	0.0005		0.00109***
	(2.70)	(3.40)	(1.56)		(4.07)
$\boldsymbol{\rho}$	−0.0008	0.0001***	−0.0006	*	0.0002***
	(−0.23)	(3.32)	(−1.87)		(4.40)
DG			0.0189***		−0.0176***
			(13.04)		(−11.12)
CAP			0.0188***		−0.0043***
			(17.86)		(−3.91)
N	63,364	63,364	63,364		63,364

volatility of cash flows for banks in our sample. These results are statistically significant at the 1% level or better. The sign on the control variables is as expected – an increase in deposits and bank capital is associated with higher average profitability and lower standard deviation of profitability over the same 5-year period.

Impact of diversification on bank credit supply

Do diversified banks lend more? We test the association between diversification and bank credit supply through the following panel regression framework:

$$\begin{aligned} CREDIT_{j,t} &= \text{const.} + TFE + \beta^{PR} NonCore + \beta^{\rho} \boldsymbol{\rho}_{j,t} + \beta^{DG} DG_{j,t} \\ &+ \beta^{CAP} CAP_{j,t} + \beta^{PA} P_{j,t} \beta^{PS} \sigma P_{j,t} + \epsilon_t^{CREDIT} \end{aligned} \tag{3.2}$$

Here, $CREDIT_{j,t}$ is one of the four proxies for credit supply for bank j in quarter t: total credit, commercial credit, real estate credit, and credit commitments. As described above, all proxies for credit supply are scaled by beginning-of-period book value of assets, and are

averaged over the same 5-year period over which we measure coinsurance and the ratio of non-core to total income.Given the evidence above that overall profitability is sensitive to our measures of diversification, we also include the level and volatility of total income to capture the effect of diversification on bank's operations that extents beyond profitability.

Estimates of the model in equation 3.2 are presented below. Each column refers to a different proxy of credit. Results are reported without (first four columns) and with (last four columns) control variables. We see that the extent of diversification has a positive impact on a bank's credit supply. As the proportion of non-core to total income increases, credit supplied to customers increases. The results also show that an increase in coinsurance (i.e. a decrease in $\boldsymbol{\rho}$) is also robustly associated with an increase in credit. In particular, a 1% increase in coinsurance is accompanied with a 2.66% increase in credit supply. Note that overall credit for all banks in our sample stands at $7.06 trillion at the end of 2014. A 2.66% increase implies a $1.88 billion increase in quarterly total loans. Thus the effect is economically quite large.

Higher coinsurance is associated not only with higher total credit supply but also higher commercial, real estate, and commitment credit. The coefficients are statistically significant at the 5% level or better. The fact that the relationship between diversification/coinsurance and average credit holds over a 5-year period is significant as it indicates that banks with lower degree of diversification and coinsurance do not simply delay credit supply to match their cash flow realizations, but that they may forego opportunities to supply credit altogether. It is also noteworthy that the inclusion of the control variables has little impact on the effect of the diversification and coinsurance measure.

Impact of diversification on financial constraints of banks

We conduct two complementary tests to investigate the impact of diversification on bank's financial constraints. First, we check if banks with more diversification pay more dividends.

We report the estimates of the following panel regression:

$$DIV_{j,t} = \text{const.} + TFE + \beta^{PR} NonCore_{j,t} + \beta^{\rho} \boldsymbol{\rho}_{j,t} + Controls_{j,t} + \epsilon_{j,t}^{DIV} \tag{3.3}$$

Here, $DIV_{j,t}$ denotes dividend payout (as a ratio of total income) for bank j in quarter t, averaged over the 5-year period over which we measure $\boldsymbol{\rho}_{j,t}$ and $NonCore_{j,t}$. In accordance with if banks with lower diversification/coinsurance are financially constrained, they should store liquidity, and therefore pay lower dividends. In other words, we expect β^{PR} to be positive, and β^{ρ} to be negative. The results show that while dividends paid to banks' shareholders increase in the proportion of non-core income, the relationship between coinsurance and dividends is not statistically significant. A 1% increase in the proportion of non-core income increases dividends by 0.35% and the effect is statistically significant at the 1% level. In contrast, once we control for the proportion of the non-core income, coinsurance does not have a statistically significant effect on dividends paid.

In our second test we check how diversification affects the sensitivity of bank credit supply to overall bank cash flow levels. If external finance is costly, banks may become financially constrained when internal cash flow levels are insufficient. Under these circumstances, bank credit supply will be sensitive to levels of cash flow generated internally. If coinsurance helps relieve financial constraints, we expect the sensitivity of bank credit to its internal cash flows to decrease in the degree of coinsurance. Thus, we estimate the following interacted model:

$$\begin{aligned} CREDIT_{j,t} &= \text{const.} + TFE + \beta^{PR} NonCore_{j,t} + \beta^{\rho} \boldsymbol{\rho}_{j,t} + Controls_t \\ &+ \beta^{PA} P_{j,t} + \beta^{I1} NonCore_{j,t} \times P_{j,t} + \beta^{I2} \boldsymbol{\rho}_{j,t} \times P_{j,t} + \epsilon_{j,t}^{CREDIT} \end{aligned} \tag{3.4}$$

Estimates of the model are presented in Table 3.5. The effect of profitability on credit supply (β^{PA}) is positive, indicating that in markets with costly external finance bank credit supply depends on the availability of internal cash flows. However, the sensitivity of bank credit supply to internal cash flows is lower for banks with a higher proportion of non-core

Table 3.5: Impact of diversification on dividends paid by banks

Notes: This table presents results of the following panel regression:

$$DIV_{j,t} = \text{const.} + TFE + \beta^{PR} NonCore_{j,t} + \beta^{\rho} \boldsymbol{\rho}_{j,t} + Controls_t + \epsilon^{DIV}_{j,t}$$

Here, $DIV_{j,t}$ denotes dividend payout (as a ratio of total income) for bank j in quarter t, averaged over the 5-year period over which we measure $\boldsymbol{\rho}_{j,t}$ and $NonCore_{j,t}$. DG, and CAP are respectively the quarterly growth rate of deposits, and the ratio of book value of equity to total book value of assets, averaged over the last 5 years. TFE are time fixed effects. Statistical significance at the 10%, 5% and 1% levels is denoted by *, **, and *** respectively. In parentheses we report t-statistics based on standard errors clustered by bank. Quarterly data, September 1987 to December 2013.

	(1)	(2)	(3)	(4)
$NonCore$	0.3480^{***}	0.3530^{***}	0.3040^{***}	0.3250^{***}
	(3.58)	(3.64)	(3.19)	(3.44)
$\boldsymbol{\rho}$	0.0056	0.0144	0.0068	0.0147
	(0.83)	(1.40)	(0.66)	(1.45)
DG		-3.1520^{***}		-3.5680^{***}
		(-8.11)		(-9.18)
CAP		1.121^{***}		0.5680^{*}
		(3.99)		(1.92)
P			32.6700^{***}	31.9400^{***}
			(5.55)	(5.15)
σP			-16.6200^{***}	-10.78^{**}
			(-3.10)	(-1.99)
N	63,364	63,364	63,364	63,364

Table 3.6: Impact of coinsurance on sensitivity of credit to internal cash flows

Notes: This table presents results of the following panel regression:

$$\begin{aligned} CREDIT_{j,t} &= \text{const.} + TFE + \beta^{PR} NonCore_{j,t} + \beta^{\rho} \boldsymbol{\rho}_{j,t} + Controls_t \\ &+ \beta^{PA} P_{j,t} + \beta^{I1} NonCore_{j,t} \times P_{j,t} + \beta^{I2} \boldsymbol{\rho}_{j,t} \times P_{j,t} + \epsilon_{j,t}^{CREDIT} \end{aligned}$$

Here, $CREDIT$ is either the total credit, commercial credit, real estate credit, or credit commitments for bank j scaled by the beginning-of-period book value of assets. The ratio of non-core to core income ($NonCore$) and their correlation $\boldsymbol{\rho}$ are computed over the last 5 years. $P_{j,t}$ denotes total income scaled by total assets for bank j in quarter t, averaged over the 5-year period over which we measure $\boldsymbol{\rho}_{j,t}$ and $NonCore_{j,t}$. The standard deviation of cash flow, $\sigma P_{j,t}$, is also computed over the same period. DG, and CAP are respectively the quarterly growth rate of deposits, and the ratio of book value of equity to total book value of assets, averaged over the last 5 years. TFE are time fixed effects. Statistical significance at the 10%, 5% and 1% levels is denoted by *, **, and *** respectively. In parentheses we report t-statistics based on standard errors clustered by bank. Quarterly data, September 1987 to December 2014.

	Total		Commercial		Real		Commit	Total		Commercial
$NonCore$	0.1550***		0.0857***		0.4500***		0.1690***	0.1410***		0.0856***
	(2.95)		(3.22)		(8.31)		(6.01)	(2.55)		(3.13)
$\boldsymbol{\rho}$	−0.0161	**	−0.0118***		−0.0053		−0.0025	−0.0150	**	−0.0117***
	(−2.24)		(−2.62)		(−0.66)		(−0.59)	(−2.09)		(−2.61)
DG								1.0900***		0.4690***
								(7.39)		(5.25)
CAP								−0.5690***		−0.0851
								(−4.94)		(−1.32)
P	10.5100***		2.1320		14.4100***		0.4230	8.1790***		2.3380
	(3.84)		(1.26)		(5.13)		(0.32)	(2.90)		(1.32)
$NonCore \times P$	−2.4570		−1.9890		−2.7790	*	−9.1920	−1.9320		−1.4530
	(−1.45)		(−0.21)		(−1.87)		(−0.94)	(−1.07)		(−0.15)
$\boldsymbol{\rho} \times P$	3.9320		2.0510		6.9440	**	0.4800	5.0720	*	1.6510
	(1.49)		(1.23)		(2.41)		(0.33)	(1.90)		(0.98)
N	63,364		63,364		63,364		63,364	63,364		63,364

income (that is, β^{I1} is negative). A similar reasoning (with opposite signs) applies to the correlation measure $\boldsymbol{\rho}_{j,t}$: less diversified income stream as well as lower coinsurance increase banks' financial constraints, thereby making bank credit supply more sensitive to internally generated cash flows (that is, β^{I2} is positive). In sum, the sign on the interaction terms are consistent with the prior that diversification relaxes financial constraints. We conclude that bank credit supply is not only sensitive to internal cash flows, but this sensitivity decreases in diversification/coinsurance.

Impact of diversification on bank risk exposures

We analyze how a bank's diversification into non-core activities impacts its equity idiosyncratic and systematic risk. Our prior is that diversification reduces a bank's idiosyncratic risk. However, the effect of diversification on bank's systematic risk is empirically an open

question.

We begin by matching the Call Report Data for all *publicly-listed* banks in our sample to monthly data for their equity returns available from the Center for Research in Security Prices (CRSP).[10] We build portfolios of bank stocks using the standard portfolio formation strategy of Fama and French (1993). In each month, we rank banks in quintiles based on the extent of bank's diversification into non-core activities, measured alternatively by (a) the ratio of non-core to total income, and (b) the degree of coinsurance i.e. $\boldsymbol{\rho}$. We track the bank return in any given quintile in the subsequent month, and calculate value-weighted returns for each portfolio. This procedure results in monthly value-weighted returns for five diversification-sorted portfolios.

To obtain systematic risk exposures, we follow a standard performance-attribution approach. We regress returns to each diversification-sorted portfolio j in excess of the one-month risk-free rate on a set of risk factors f_t:

$$r_{j,t} - r_{f,t} = \alpha_j + \beta_j f_t + \epsilon_{j,t} \tag{3.5}$$

The vector f_t contains six factors that have been previously documented to capture differences in average returns to nonfinancial stocks and bonds, $f_t = [market\ smb\ hml\ ltg\ crd\ liq]$. The variables *market*, *smb*, and *hml* represent the returns on the three Fama French stock factors. The factors are constructed using the six value-weighted portfolios of all stocks on NYSE, Amex and NASDAQ (including financials) formed on size and book-to-market. We capture market excess returns using the value-weighted return on all NYSE, Amex and NASDAQ stocks (from CRSP) minus the one-month Treasury bill rate (from Ibbotson Associates). A bank manages a portfolio of bonds of varying maturities and credit risk. Therefore, we augment the model with two additional bond factors: *ltg*, the excess returns on an index of 10-year bonds issued by the U.S. Treasury (from Global Financial Data); and *crd*, the

[10]For matching, we use the link file maintained by the Federal Reserve Bank of New York containing the CRSP permanent number codes for publicly-listed banks.

excess returns on an index of investment grade corporate bonds (maintained by Dow Jones). Finally, we include the Pastor and Stambaugh (2003) liquidity risk factor, denoted liq.

For brevity, we show loadings only on market risk factors and the adjusted R^2. Each panel refers to a different measure of diversification. We begin by looking at Panel A, where portfolios are sorted by the proportion of non-core income. Portfolio 1 contains the banks with the lowest proportion of non-core income. We note that the adjusted R^2 increases monotonically from the first quintile to the last quintile. That is, the idiosyncratic risk (or, $1-R^2$) of banks that derive a higher portion of their income from non-core activities is lower than the idiosyncratic risk of other banks. This evidence is consistent with the phenomenal rise of securitization-driven lending over our sample period. The typical securitization bundle was structured to diversify away local risk by combining mortgages originated in different parts of the country. If banks that increased the proportion of income from non-core activities did so by relying more on securitization activity, this would result into a reduction of idiosyncratic risks, as it does in Table 5

The table also reports the level of total risk of monthly stock returns to diversification-sorted portfolios, as measured by their sample standard deviation. Clearly, total riskiness of banks' equity claims increases in the proportion of income derived from non-core activities. One possible reason for this result is the fact that non-core income includes income from hedge funds, venture funds, private equity funds, and trading and income from these activities are generally more volatile. Thus, despite the secular decline in the total riskiness of the bank sector over our sample period, banks that increasingly rely on non-core income boost their total risk exposures compared to others.

To study the dynamics of systematic risk of banks we look at the betas on the value-weighted market index, that is, their CAPM betas. Market beta increases monotonically with the proportion of non-core income. Over the entire sample, banks with the largest proportion of non-core income have a market beta of 1.31 as compared to 0.75 for those with the lowest proportion of non-core income. It is also important to note that the increase in

market betas from group 1 to group 5 cannot be explained by balance sheet leverage ratios. The average ratio of capital to total assets for banks in group 1 is 10.11%, which is very close to that for banks in group 5 at 10.76%. Overall, these findings show that banks that rely more on non-core activities have lower idiosyncratic risk, but higher total risk.

Finally, motivated by the evidence from the correlations, that the coinsurance provided by diversified subsidiaries changes over time, we explore one possible reason for why diversification increases systematic as well as total risk. The last row of Panel A presents the loading on the correlation risk factor, which is particularly suited to capture risk exposure to unexpected changes in correlation.[11] There is mounting evidence that asset return correlations change over time, and tend to increase at times of aggregate economic uncertainty. These are precisely the states of nature during which cash flow correlations of banks are also likely to change. For this reason, we expect the market-based correlation risk measure to be a good proxy for the cash-flow correlation risk faced by banks. Further, we have shown that banks derive a significant portion of their non-core income from trading, venture capital, and hedge funds activities sponsored by banks. These activities allow banks to trade in options and derivatives and this exposes them directly to market-based correlation risk. We observe that the loading on the correlation risk factor decreases monotonically in the proportion of non-core income. In particular, the exposure almost doubles when moving for the lowest (coefficient of -12.97) to the highest (coefficient of -21.94) diversification-sorted portfolio. Because of a negative correlation risk premium, so more negative coefficients reveal higher exposure to correlation risk. Thus, banks which derive a higher proportion of their income from non-core activities appear to be more exposed.

Are bank managers aware that diversification increases systematic as well as correlation risk? If yes, why do they still choose to diversify bank activities? One possible answer is that diversification allows banks to grow rapidly beyond a specific size threshold. A large literature shows that banks that exceed a particular threshold benefit from implicit and

[11]We are grateful to the authors for sharing their data with us.

explicit guarantees provided by regulators. For example, Gandhi and Lustig (2015) find large banks have a lower cost of capital due to implicit government guarantees. Thus, it is likely that managers diversify to receive benefits of implicit guarantee and this exceeds the costs of diversification outlined in this section.

To test if this is the case, we regress returns of diversification-sorted portfolios of banks on the size factor of Gandhi and Lustig (2015), who show that banks that benefit from an implicit government guarantee have a more negative loading on the size factor. The last row in panel B reports the loadings on the size factor for diversification-sorted portfolios of banks. We see that the loading on the size factor increases near monotonically from portfolio '1' to portfolio '5'. Some of this relationship is clearly mechanical as size and diversification go together. However, the fact that diversified banks benefit more from implicit government guarantees is consistent with the story that one motivation for banks to diversify is to grow rapidly beyond a certain size threshold to benefit from implicit government guarantees.

Panel B reports the results for portfolios of banks that are now sorted by the degree of coinsurance as measured by $\boldsymbol{\rho}$. Banks with the lowest coinsurance (and hence most diversified) are in portfolio 1. We find that although banks sorted by $\boldsymbol{\rho}$ also have a significant exposure to correlation risk, neither the correlation risk exposure nor the proportion of systematic risk is monotone in $\boldsymbol{\rho}$. Returns to banks in portfolio 1 have a standard deviation of 24%, slightly higher than that of banks in portfolio 5 which stand at about 22%. However, as expected, systematic risk accounts for a lower proportion of total risk for banks in portfolio 1 as compared to those in portfolio 5.

To sum up, as banks increase their reliance on non-core activities, they reduce their idiosyncratic risk exposures. This pattern is consistent with the argument in models of financial intermediation that banks' managers try to diversify idiosyncratic risk to minimize the cost of delegated monitoring. It also indicates why better diversification is associated with a higher levels of credit supply, profitability, and lower financial constraints for banks. However, reliance on non-core activities increases total risk as well as exposure to systematic

Table 3.7: Risk profile of diversification-sorted portfolios of banks

Notes: This table presents the estimates from the OLS regression of monthly excess returns on diversification-sorted portfolio of banks on the following six risk factors: *market*, *smb*, and *hml* denote the three Fama-French stock factors; *ltg* denotes the excess return on an index of long-term government bonds; *crd* denotes the excess return on an index of investment-grade corporate bonds; *liq* denotes the liquidity risk factor of Pastor and Stambaugh (2003). Results are presented for portfolios of banks that are sorted in quintiles based on alternatively the proportion of non-core income (Panel A) or the degree of coinsurance (Panel B). We report the loading on *market*, the adjusted R^2, and the sample standard deviation (σ). We also report the loading on the correlation risk factor (*corr*), and on the size factor of Gandhi and Lustig ($size\,factor$). Statistical significance at the 10%, 5% and 1% levels is denoted by *, **, and *** respectively. In parentheses we report t-statistics based on Newey-West standard errors with 3 lags. Monthly data, September 1987 to December 2014.

Portfolio	1	2	3	4	5
		Panel A : Proportion of non-core income			
market	0.7499***	0.7616***	0.7190***	1.0218***	1.3117***
	(10.20)	(12.57)	(12.18)	(15.60)	(20.38)
R^2(%)	47.66	50.79	48.73	59.02	67.03
σ(%)	20.38	19.50	18.81	23.20	28.29
corr	−0.1252***	−0.1293 *	−0.1621 *	−0.1815***	−0.1990***
	(−4.31)	(−5.01)	(−4.76)	(−5.01)	(−4.75)
$size\,factor$	−0.3286***	−0.2968 *	−0.3802 *	−0.4137***	−0.4733***
	(−2.88)	(−4.66)	(−4.84)	(−3.68)	(−5.55)
		Panel B : Degree of coinsurance			
market	1.1200***	1.1400***	0.9900***	1.3900***	1.1800***
	(14.50)	(16.15)	(14.13)	(16.92)	(20.58)
R^2	51.89	54.64	50.04	53.99	66.64
σ(%)	24.15	22.82	21.57	28.02	21.60
corr	−0.2087***	−0.1706***	−0.1587***	−0.2192***	−0.1676***
	(−5.07)	(−4.40)	(−4.31)	(−4.43)	(−4.47)
$size\,factor$	−0.3690***	−0.3619***	−0.3671***	−0.4333***	−0.3793***
	(−9.16)	(−8.72)	(−8.66)	(−8.78)	(−12.04)

risk and in particular correlation risk. Correlation risk arises because of unexpected changes in the relationship between incomes from diversified subsidiaries, which typically occur during periods when the benefits of diversification are most needed (i.e. high marginal utility states). Despite these disadvantages to diversification, bank managers may consider diversification optimal if it allows their banks to grow large in order to benefit from implicit and explicit government guarantees. Finally we see that for banks that are careful to design non-core income to diversify cash flows, total risk is not much higher than other banks and also systematic risk accounts for a lower proportion of systematic risk.

Impact of diversification on sensitivity to macroeconomic conditions

The exposure of diversified banks to shocks in correlation, which typically occur in bad times, suggests that these banks' operations ought to be more sensitive to changes in the investment opportunity set. That is, banks with different degree of diversification are expected to react differently to changes in macroeconomic conditions. We examine this prediction by re-estimating our baseline panel regression of credit supply where we now replace the time fixed effects by the change in GDP growth in the previous quarter. The model is estimated separately for banks with a high proportion of non-core income (banks in portfolio 5, Panel A) and for banks with a low proportion of non-core income (banks in portfolio 1, Panel B).

We find that for banks with a higher proportion of non-core income, a 1% decrease in GDP (i.e. worsening business conditions) leads to a 2.177% decrease in total credit. For banks with a low proportion of non-core income, the magnitude of this effect is much smaller as a 1% decrease in GDP decreases credit by a modest 0.907%. In sum, the same aggregate shock has a much more pronounced effect (nearly 2.5 times higher) on banks that rely more heavily on non-core income compared to less diversified banks. This finding lends further support to our claim that while diversification reduces idiosyncratic risk, it increases banks' exposure to systematic risk.

The evidence that, for diversified banks, credit supply is more sensitive to macroeconomic conditions may appear in contradiction with the fact that these banks on average also supply more credit. One potential explanation is that the benefits of diversifying into non-core activities are limited to "good times", that is, periods of economic expansion. To verify whether this is indeed the case, we rerun our earlier credit supply analysis, but now interact both the proportion on non-core income and the correlation with a dummy variable that equals one for all those quarters in which the U.S. economy is an economic contraction (i.e. a recession). This allows us to analyze how diversification aids (or hinders) bank credit supply in and outside periods of aggregate shocks. Results are presented below.

The table shows that the beneficial effects of diversification are limited to periods in which

Table 3.8: Impact of diversification on bank credit supply in and outside recessions

Notes: This table presents results of the following panel regression:

$$CREDIT_{j,t} = \text{const.} + TFE + \beta^{PR} NonCore_{j,t} + \beta^{\rho} \boldsymbol{\rho}_{j,t} + \beta^{I1} NonCore_{j,t} \times D + \beta^{\rho} \boldsymbol{\rho}_{j,t} \times D + \beta^{DG} DG_{j,t} + \beta^{CAP} CAP_{j,t} + \epsilon_{j,t}^{CREDIT}$$

Here, $CREDIT$ is either the total credit, commercial credit, real estate credit, or credit commitments for bank j scaled by the beginning-of-period book value of assets. The ratio of non-core to core income ($NonCore$) and their correlation $\boldsymbol{\rho}$ are computed over the last 5 years. $CREDIT$, DG, and CAP are respectively credit supply, the quarterly growth rate of deposits, and the ratio of book value of equity to total book value of assets, averaged over the last 5 years. D is a dummy variable that equals TFE are time fixed effects. Statistical significance at the 10%, 5% and 1% levels is denoted by *, **, and *** respectively. In parentheses we report t-statistics based on standard errors clustered by bank. Quarterly data, September 1987 to December 2014.

	Total	Commercial	Real	Commit	Total	Commercial
NC	0.0797^{**}	0.0953^{***}	0.3760^{***}	0.1500^{***}	0.0792^{*}	0.0934^{***}
	(2.08)	(4.90)	(10.65)	(7.55)	(1.97)	(4.66)
$\boldsymbol{\rho}$	-0.0263^{***}	-0.0061^{***}	-0.0234^{***}	-0.0008	-0.0284^{***}	-0.0071^{***}
	(-6.89)	(-2.71)	(-5.71)	(-0.50)	(-7.57)	(-3.16)
$NC \times D$	-0.1180^{***}	-0.0479^{***}	-0.0543^{**}	-0.0688^{***}	-0.0951^{***}	-0.0406^{***}
	(-4.35)	(-3.41)	(-2.04)	(-4.04)	(-3.44)	(-2.77)
$\boldsymbol{\rho} \times D$	0.0028	-0.0046	0.0050	-0.0030	0.0084	-0.0026
	(0.43)	(-1.19)	(0.68)	(-0.76)	(1.30)	(-0.69)
DG					0.9730^{***}	0.4280^{***}
					(6.85)	(5.00)
CAP					-0.6800^{***}	-0.1250^{**}
					(-6.24)	(-2.07)
N	63,364	63,364	63,364	63,364	63,364	63,364

the U.S. economy does not face an aggregate economic shock. While banks with a higher proportion of non-core income on average supply more credit, in recessions they supply less credit as compared to less-diversified banks. Similarly, the beneficial effects of correlation are also limited to periods in which the U.S. economy is not in recession. In recessions, coinsurance is not positively associated with a higher credit supply and the effect is not not statistically significant in any specification. The fact that the U.S. economy has just spent 10% of the time in recessions during our sample period, explains why diversification is on average associated with higher credit and profitability.

Thus, we conclude that the beneficial effects of diversification are limited to good times, when perhaps financial constraints on banks are not likely to bind in any case. The evidence is consistent with the analysis in the previous section that shows diversified banks are more exposed to correlation risk. This exposure likely increases the cost of capital of diversified banks in recessions, and as a result providing credit is costly (or uncompetitive) for such banks.

3.5 Robustness tests

A potential concern with our methodology is reverse causality or endogeneity among some of our key variables. For example, consider the relationship between profitability and diversification. On one hand, existing agency cost models suggest that managers of more profitable banks, with cash flow in excess of that required to fund positive NPV projects, are likely to invest such cash flow inefficiently, to grow banks beyond optimal size and scope. On the other hand, existing risk management models suggest that diversification can help improve profitability, for example, by helping banks avoid deadweight financial costs.

Thus, profitability can be a determinant (due to agency costs) or the result (due to effective risk management) of banks' decision to diversify. However, there are crucial differences. While agency theory suggests that *past* profitability impacts banks' decision to diversify, risk

management theory suggests that diversification impacts *future* profitability. In addition, while agency theory implies a *negative* relationship between diversification decisions and *future* profitability, risk management implies a *positive* relationship between these variables.

As another example, consider the relationship between growth rates and diversification. On one hand, banks with high growth rates have a high probability of facing financing shortfalls, and will benefit most from diversification. On the other hand, diversification synergies result in lower volatility of cash flows and higher growth rates. Thus, (as is the case with profitability above) growth rates can both be a cause and effect of diversification. However, again there are crucial differences in the timing and the direction of the effect. One strand of the literature relates past growth rates to diversification, whereas the other strand relates diversification to future growth rates. Further, bank diversification should positively impact future growth rates, only if it's value enhancing.

In light of this discussion, in our analysis, we are careful in distinguishing between the impact of ex-ante and ex-post variables on banks' decision to diversify into non-core activities. In addition, in this section we carry out a battery of robustness checks to dispel endogeneity concerns.

3.5.1 Difference-in-difference test

We examine what effect, if any, does initiation of banks' participation in non-core activities have on their core intermediation capabilities (i.e. credit supply). Unfortunately, none of the banks in our sample "initiate" participation in non-core income. However, some banks do begin participating in trading activities. The correlations show that (like other non-core activities) trading income is also negatively correlated with core income and therefore also provides coinsurance benefits. In addition, since trading is one of the largest and most controversial component of non-core income, the effect of trading activities on banks' core intermediation capabilities is of independent interest.

We identify banks that initiate trading activities as follows. In any given quarter t, a

bank is said to be participating in trading activities if it either has non-zero trading income or non-zero trading assets. In the same quarter, a bank with zero trading assets and income may still be classified as participating in trading activities, if it was classified as such in any one of the 4 previous quarters (i.e. from $t-4$ to t).

We consider the period that spans from four quarters before to four quarters after a bank begins participating in trading activities using a difference-in-difference approach. This approach allows us to compare credit supply policy of banks that begin participating in trading activities with their credit supply policy right before initiation, and to other banks that are simultaneously not participating in trading activities. Specifically, we estimate the following regression:

$$CREDIT_{j,t} = \text{const.} + \sum_{i=4}^{i=-4} \theta_i + DG_{j,t} + CAP_{j,t} + \epsilon_t^{CREDIT} \tag{3.6}$$

Here, the dummy variables θ_i take values 1 in the i^{th} quarter before or after a bank starts trading, and is zero otherwise. Negative values of i refer to quarters before a bank starts trading. The quarter in which a bank initiates trading activities is normalized to $i = 0$. We plot the estimates for these dummy variables θ_i. Each panel refers to a separate proxy of credit supply. The top left panel presents the results for total credit. The figure shows how initiation of trading activities impacts the average supply of credit for banks in our sample. Average credit supply increases sharply after a bank begins participation in trading activities. The difference is statistically significant, that is, the confidence intervals on θ_i measured before and after initiation of trading activities do not overlap. The effect is most pronounced for real estate credit and credit commitments. The fact that banks lend more via commitments (after initiation of trading trading activities) is consistent not only with an improvement in current liquidity constraints, but also with an increase in banks' confidence in managing liquidity constraints in the future. This likely makes them more comfortable lending via credit commitments.

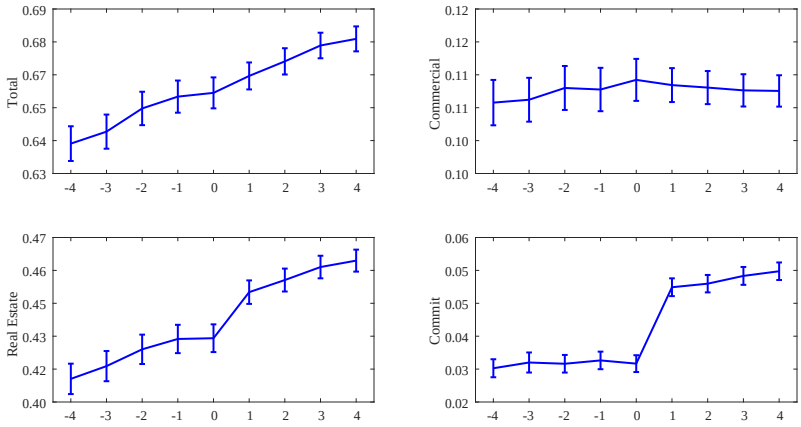


Figure 3.2: Average credit supply before and after banks start trading

This figure plots the average credit supply for banks in our sample before and after initiation of trading activities. We normalize the quarter in which each bank in our sample starts trading to 0. A bank is classified to be participating in trading activities if it has non-zero trading income or non-zero trading assets. In any quarter, a bank with zero trading assets and trading income may still be classified as participating in trading activities if it was classified as such in any one of the 4 previous quarters. Each panel plots the values of the dummy variables from the diff-in-diff specification for a different proxy of credit. The bar indicates plus and minus two standard error bands. Quarterly data, September 1987 to December 2014.

While the difference-in-difference exercise is the closest we can get in terms of establishing causality, it does not obviously alleviate the issue of endogeneity in the decision to participate in trading activities. Also, the analysis can only be applied to the decision of a bank to participate in trading activities, as the extent of coinsurance provided by trading activities is not a binary variable and cannot be taken into account.

The results of the difference-in-difference should not be viewed as a comprehensive assessment of the costs and benefits of bank's participation in trading activities. These results also do not imply that all banks should participate in trading activities. The beneficial effects of scope-diversification are limited to periods in which the U.S. economy is not in recession. Our results do suggest that, in taking the decision whether to engage in such activities managers should balance the benefits with the implicit costs we outlined above.

3.5.2 Endogeneity concerns and additional robustness checks

In all our previous tests, correlation and average credit are measured over the same 5-year period. This may bias our inference in the presence of endogeneity. Correlation could be

driven by lending, and not viceversa. In addition, as argued above, there is the issue of reverse causality. Ideally, we would like to observe exogenous shocks to a variable correlated with our coinsurance statistic. Unfortunately, there are no opportunities to exploit this approach in our data.

Below, we run different specifications of our main model which all aim at ameliorating concerns regarding endogneity and robustness. In column titled 'Baseline', the dependent variable, i.e. credit supply, is computed using data one quarter after the computation of $NonCore$ and $\boldsymbol{\rho}$. In other words, $NonCore$ and $\boldsymbol{\rho}$ is estimated using backward-looking 5-year data over $t-20$ to t, while the left-hand-side variables are measured only using data in quarter $t+1$. The control variables (deposit growth and capital) are also measured at t, and are not averaged over $t-20$ to t. In the column entitled 'Non-overlap', we report the results for a deliberately conservative approach, wherein we estimate diversification measures on non-overlapping 5-year windows. This check addresses concerns about persistence and proper computation of standard errors on overlapping series. The column titled '3-year' use correlation estimated over 3-year window (i.e. from $t-12$ to t) instead of the 5-year correlations that are used in all our tests. Finally, the column titled 'Covariance' uses the covariance as a measure of coinsurance between banks' core and non-core income, thereby simultaneously taking into account the size and correlation of the income generated by these activities. Given our previous findings, in all specifications we include a dummy variable D that equals one whenever the U.S. economy is a recession as classified by the NBER. The table shows that our central conclusions remain unchanged. Diversification into non-core activities helps banks but the benefits are limited to periods in which the U.S. economy does not face aggregate shocks.

Table 3.9: Robustness checks

Notes: This table presents the results for different robustness checks. In the first column, we estimate equation Panel Credit when the dependent variables are computed in quarter $t+1$ only and the coinsurance ρ is measured over the period from $t-20$ to t. In the second column, we estimate equation Panel Credit on non-overlapping 5-year periods. In the third column, we use data for lagged 3 years (instead of the usual 5 years) to estimate our diversification measures. In the last column, we estimate equation Panel Credit but now measure coinsurance by covariance, rather than correlation, between core and non-core incomes over the last 5 years. Statistical significance at the 10%, 5% and 1% levels is denoted by *, **, and *** respectively. In parentheses we report t-statistics based on standard errors clustered by bank. Quarterly data, September 1987 to December 2014.

	Baseline	*Non-overlap*	3-*year*	*Covariance*
$NonCore$	0.0833^{**}	0.0443	0.0658^{**}	0.0789^{*}
	2.0400	1.0700	1.9700	1.9400
$\boldsymbol{\rho}$	-0.0281^{***}	-0.0287^{***}	-0.0219^{***}	-2.5330^{***}
	-7.3000	-6.2800	-8.6600	-3.5100
$NonCore \times D$	-0.0902^{***}	-0.1300^{*}	-0.1090^{***}	-0.1160^{***}
	-3.6000	-1.8400	-4.0500	-4.2600
$\boldsymbol{\rho} \times D$	0.0023	-0.0059	0.0129^{***}	-0.2410
	0.3500	-0.4600	2.6900	-0.1300
DG	0.9530^{***}	1.0370^{***}	0.7850^{***}	0.9100^{***}
	6.5300	7.2000	8.2700	6.3100
CAP	-0.6800^{***}	-0.6740^{***}	-0.6600^{***}	-0.7070^{***}
	-6.1200	-6.4400	-7.3500	-6.4000
N	$61,302$	$4,021$	$81,550$	$63,364$

3.6 Conclusion

Over 1987 to 2014, U.S. banks have increasingly diversified into non-core activities such that income from these activities now accounts for a majority of bank sector income. This behavior appears to run counter to existing corporate finance literature that suggests diversification destroys firm value, but is consistent with extant models of financial intermediation that suggest banks diversify to reduce idiosyncratic risk and achieve credibility as monitors of borrowers. Indeed, as predicted by this literature, large banks, banks with high distress costs, and banks facing high cost of raising external finance diversify more aggressively.

For the average bank, non-core income offsets risks elsewhere in its balance sheet. That is, the average bank that engages in non-core activities benefits from coinsurance on account of the imperfect correlation between cash flows from core and non-core activities. Better diversification/coinsurance for banks is associated with higher profitability, higher average credit supply, and lower financial constraints.

However, diversification has not translated into real reductions in risk. For one, the benefits of diversification/coinsurance listed above are limited to good times (periods when the U.S. economy is not in recessions), when such benefits are needed least and financial constraints are likely not binding. In addition, we find that, for publicly listed banks, diversification increases their exposure to correlation risk. Correlation risk arises because of an unexpected change in the relationship between core and non-core incomes, which can be linked to an adverse evolution of diversification opportunities. This is precisely what happened during the credit crisis of 2007 to 2009.

Future researchers can address important questions related to our results. In particular, what actually causes the degree of coinsurance to vary across banks and how this may relate to the costs of participating in non-core activities are questions we have not yet fully answered.

In the wake of the financial crises, a regulatory debate is now centered on the optimal scope of banking activities. The Volcker Rule advocates the segregation of some activities

from traditional banking operations. Similar rules to compel banks to divest their non-core activities have been proposed in the United Kingdom and several other countries. These proposals are opposed by banks on the grounds that diversification helps banks diversify cash flows and manage risks in a way that improves overall bank sector safety. Our results indicate that benefits to diversification may be limited, providing additional information to academics, regulators and practitioners to assess the costs and benefits of participating in non-core activities.

We collect balance sheet data for banks from the 'Report for Condition and Income' (henceforth Call Reports) required to be filed by all FDIC-insured bank holding companies in the U.S. This data is available at https://www.chicagofed.org/applications/bhc_data/bhcdata_index.cfm. Definitions for the variables are available at http://www.federalreserve.gov/apps/mdrm/. Banks with total book value of assets above $500 million file this report quarterly. Other banks file this report only semi-annually. We restrict our sample to banks which file the Call Reports quarterly and report a positive book value of assets. Between September 1987 and December 2014, this yields 160,761 observations. The actual number of observations in our analysis is less for several reasons. First, we require that the banks in our sample have five consecutive years (20 quarters) of data continuously available. This leaves us with 128,953 bank-quarters. Next, in order to make sure that outliers are not driving our results, we eliminate any observations in which the quarterly growth rate in the total book value of assets is more than three standard deviations from its mean. This leaves us with 123,487 observations. Finally, since our regressions involve correlations computed over 20 quarters (5 years), we lose the initial 20 observations for each bank in our sample so that our total sample size is 60,335.

The data present a number of challenges in terms of creating a consistent time-series. Due to changing reporting requirements, some of the data items in the Call Reports used for the construction of key variables in our analysis are not comparable across quarters. The Chicago Federal Reserve Bank provides instructions for the construction of consistent

time-series for the data in the Call Report. These instructions are available at http://www.chicagofed.org/webpages/banking/financial_institution_reports/bhc_data.cfm.

Once we define time-series for individual banks, we also compute data for all U.S. banks (i.e. the aggregate U.S. bank sector) to report summary statistics in the data section. To compute the time-series for all U.S. banks, we start with data for individual banks. We filter the top and bottom 1-percentile of banks based on the quarterly growth rate in total book value of assets. This filter removes observations for those bank-quarters in which banks are involved in significant mergers. For aggregation, we require that in each quarter, banks included in our sample have call report data available for at least 12 previous quarters (3 years). We also require that for each quarter Call Report data for a particular bank is available for the previous and current quarters. This requirement ensures that the time-series of core, non-core, and trading incomes are not affected by entry or exit of banks. This requirement also means that the actual number of banks used in any quarter to compute the time-series for all U.S. banks varies over time.

Appendices

Appendix A

Proofs of Propositions 1.2.1.

A.0.1 Technical Environment

Invariant Measure on X The *kernel* of an homogeneous Markov chain is a time-invariant function $m(x^{[i]}, x^{[j]}) : X \times 2^X \longrightarrow [0,1]$ containing the transition probabilities for X_t, $\Pr(X_t = x^{[j]} | X_{t-1} = x^{[i]}) =: m(x^{[i]}, x^{[j]})$ for every ordered pair of states. Let $S = \#X$; then $\sum_{k=1}^{S} m(x^{[i]}, x^{[k]}) = 1$ for each i. The transition matrix induced by the kernel has entries $(\mathcal{M})_{i,j} = m(x^{[i]}, x^{[j]})$.

We assume the chain is irreducible, i.e., $m(x^{[k]}, x^{[j]}) > 0$ for every $0 \leq k, j \leq S$. We also assume $\mathcal{M}$ is aperiodic: $\gcd\{n : m(x^{[j]}, x^{[j]})^n > 0\} = 1$ for each $x^{[j]} \in X$ ("gcd" finds the greatest common divisor). In finite states irreducibility and aperiodicity reduce to irreducibility for every $n \in \mathbb{N}$, $m(x^{[k]}, x^{[j]})^n > 0$, where $j = k$ captures the aperiodicity condition (cf., Hairer, 2006, E3.9 p.14).

Define the space of probability measures over X

$$F_\mu := \left\{ q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_S \end{bmatrix} : \sum_{i=1}^{S} q_i = 1,\ q_i \in [0,1] \right\} \subset \mathbb{R}^S$$

Let $\{e_i\,,\, 1 \leq i \leq S\}$ be standard basis vectors for $\mathbb{R}^S$.

Proposition A.0.1 *There is a unique invariant measure $\mu_0 \in F_\mu$ satisfying $\mu_0' = \mu_0' \mathcal{M}^N$ for every $N \geq 1$ and $e_i \cdot \mu_0 > 0$ for every $0 \leq i \leq S$.*

Proof Levin, Peres and Wilmer (2008), p. 12 Proposition 1.14

□

Consider a sequence of elements $h_n \in F_\mu$ generated by $\mathcal{M}'$ and denote the k'th outcome h_k. It is clear that when $h_0 = \mu_0$, $(\mathcal{M}')^K h_0 =: h_K = \mu_0$ for every $K \in \mathbb{N}$.

The Path Space Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define the product space $Z := \{X\}^{\mathbb{N}} = X \times X \times ... = \prod_{n\in\mathbb{N}} X_n$. Z contains all infinite sequences of elements of X. A path $z_s\, s \in \mathbb{N}$ is an arithmetic function $z : \mathbb{N} \to \mathbb{R}$. A *sample path* is a finite sequence of elements in X, $z_{s\leq N} \in \{X\}^N$ on the truncated domain $\{0, 1, 2, ..., N\}$. Events $\omega \in \Omega$ realize as paths $\omega \mapsto Z$, which we occasionally emphasize by writing $\{z\}(\omega) : \Omega \longrightarrow Z \subset \mathbb{R}^{S\mathbb{N}}$.

We use the discrete σ- field for X, written 2^X. A sequence of refinements to $(\Omega, \oslash) =: \mathcal{F}_0$, $\mathcal{F}_N \subset \mathcal{F}_{N+1}$, is *generated* by the sample-path events $\{\omega : \{z\}_N(\omega) \in \sigma(\{X\}^N)\} \in \mathcal{F}_N$ for every N. Each $\mathcal{F}_0 \subset \mathcal{F}_n \subset \mathcal{F}$ for $n < \infty$. We write $\sigma(\{X\}^N) = (2^X)^N$ for finite N.

Lemma A.0.2 $\sigma(\prod_{n\in\mathbb{N}} X_n) := \sigma(X^{\mathbb{N}}) = \sigma(X)^{\mathbb{N}}$

Proof The countable product of finite sets, 2^X, in this case given by $(2^X)^{\mathbb{N}} = \sigma(X)^{\mathbb{N}}$, is countable. Hence, using Theorem 4.44, Aliprantis and Border (2006), p.149, for countable σ- fields, $\sigma(\prod_{n\in\mathbb{N}} X_n) = \prod_{n\in\mathbb{N}} \sigma_n(X)$.

□

In words, the σ -field generated by the countable product space is equivalent to the countable product of the σ -fields generated by the state space X (which is not true in general). The implication is that any event in the sequence space can be written as the countable product of elements in 2^X, or equivalently, some countable product of elements of X.

Denote the finite-dimensional distributions of the Markov chain $\mathbb{P}_{\mu,N}(b)$, $b \in (2^X)^N$ for each initial distribution over X, $\mu(x)$. The Dirac mass δ_x on points in X is a particular initial distribution, in which case we write $\mathbb{P}_{x,\cdot}$. The probability of a particular sample path $(2^X)^N \ni z_N : x_0 \rightsquigarrow x_N$ is written

$$h_0(\{z_N\}) = \mu(x_0)m(x_0, x_1)m(x_1, x_2)...m(x_{N-1}, x_N)$$

for initial distribution μ.

The probability of any event $b_N \in (2^X)^N$ can be written as the sum of probabilities of paths $\{z_N\}$ where it is true, $\mathbb{P}_{\mu,N}(b_N) = \sum_{\{z_N\}\in b_N} h_0(\{z_N\})$, which themselves can be enumerated. To track the paths where b_N is true, each ordered configuration j in the set of configurations $J(b_N)$ is written as a path index, $x_{1(j)}, x_{2(j)}, ..., x_{N(j)}$, where the notation is shorthand for $x_{t(j)} = x_t^{k(j)}$. The set $J(b_N)$ is finite for $N < \infty$, and at most countable for $N = \infty$. The probability of any sample event is a sum of finite products of the kernel

$$\mathbb{P}_{\mu,N}(b_N) = \mu(x) \sum_{j\in J(b_N)} \prod_{n(j)=1}^{N} m(x_{n(j)-1}, x_{n(j)})$$

The preceding argument ensures the sample distributions are *consistent*,

$$\begin{aligned}\mathbb{P}_{\mu,N+1}(b_1, b_2, ..., b_N, X) &= \mu(x) \left[\sum_{j\in J(\boldsymbol{b})} \prod_{n(j)=1}^{N} m(x_{n(j)-1}, x_{n(j)})\right] \sum_{i=1}^{S} m(x_N, x_{N+1}^{[i]}) \\ &= \mathbb{P}_{\mu,N}(b_1, b_2, ..., b_N) \qquad (7.1a)\end{aligned}$$

for $b_i \in \sigma(\{X\}^i)$, every finite N and for $\boldsymbol{b} := (b_1, b_2, ..., b_N) \in (2^X) \times (2^X)^2 \times ... \times (2^X)^N$. This definition of consistency is standard (cf. Durrett (2010, p. 366)). The product measure of the chain is given by the pushforward,

$$\mathbb{P}_\mu(b) := \mathbb{P} \circ \{z\}^{-1}(b) = \mathbb{P}(\omega : \{z\}(\omega) \in b)$$

for every initial distribution μ on X.

Proposition A.0.3 *The product measure $\mathbb{P}_\mu$ exists and is uniquely characterized by the finite-dimensional distributions.*

Proof By application of the Kolmogorov extension theorem, letting $N \to \infty$ in 7.1a (Durrett (2010), p. 366, Theorem A.3.1)

□

Definition: Define the *tail* σ -field $\mathscr{T} := \bigcap_{n\geq 1}\{\sigma(\bigcup_{m\geq n} \mathcal{F}_m)\}$. An ergodic set

$$\left\{B \subset (2^X)^{\mathbb{N}} : \{\omega : \{z\}(\omega) \in B\} \subset \mathscr{T}\right\}$$

is a subset of a state space that has $\mathbb{P}_x(b) := \mathbb{P} \circ \{z\}^{-1}(b) \in \{0, 1\}$ for every $b \in B$ and each initial $x \in X$. A Markov process is ergodic if its state space is an ergodic set.

Corollary A.0.4 *An irreducible aperiodioc finite-state Markov chain is ergodic*

Proof Tuominen and Tweedie (1994), pp. 779-780, Theorem 2.3.

□

A.0.2 Decomposition and Wold Representation

Remark: A *reversible* Markov operator admits a positive real point spectral decomposition. Important departures from reversibility are of economic interest, so we avoid the assumption of reversibility. See Hansen and Scheinkman (1995) for an early discussion of irreversible Markov models in asset pricing.

Definition: The *dual* of $F(X)$ written $F(X^*)$ is the space of linear functionals η_f on $F(X)$, $\eta_f : F(X) \longrightarrow \mathbb{R}$.

Proposition A.0.5 (Decomposition)
I.
The image of X^* under $\mathcal{M}'$ is the direct sum of an invariant subspace $\mathscr{I}_0 = \mathscr{I}(\mu_0)$ and its orthogonal complement $\mathscr{I}_0^\perp$,

$$\mathscr{R}(\mathcal{M}') = \mathscr{I}_0 \bigoplus \mathscr{I}_0^\perp$$

Moreover, for any integer $k \geq 0$, the image space $\mathscr{R}([\mathcal{M}']^k)$ admits a decomposition into $\mathscr{I}_0$ and a space that depends on k

$$\mathscr{R}([\mathcal{M}']^k) = \mathscr{I}_0 \bigoplus (\mathscr{I}_0^\perp)_k$$

In particular, for $\mathcal{M}'_\gamma$: $Span(\mathcal{M}'_\gamma) = \mathscr{I}_0^\perp \subset \mathscr{R}(\mathcal{M}')$ *and* $1 \leq k \in \mathbb{N}$,

$$(\mathscr{I}_0^\perp)_k = Span\left([\mathcal{M}'_\gamma]^k\right)$$

II.
The columns of $(\mathcal{M}'_\gamma)^k, k \in \mathbb{N}$ *converge to the origin in the operator norm topology*

$$((\mathcal{M}'_\gamma)^k)_{,j} \longrightarrow_{k\longrightarrow\infty} \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

Each column $(\mathcal{M}')_{,j}$ *of* $\mathcal{M}'$ *converges*

$$(\mathcal{M}')^k_{,j} \longrightarrow \mu_0$$

in total variation to a distribution μ_0 *that is independent of* j.

We first state a lemma we will need. A general proof is provided in section 7.3.2.

Lemma A.0.6 *Recall the dual space* X^* *contains the linear functionals* η *on the range of* X *under* $\mathcal{M}$. *Then,*

- *We can identify* X^* *with the space of probability measures* $F_\mu(X)$ *over* X, $F_\mu(X) \equiv X^*$
- *The adjoint* $\mathcal{M}'$ *is a continuous automorphism on* $F_\nu(X \times X)$. *In particular, each row of* $\mathcal{M}'$ *takes* $F_\mu(X) \longrightarrow F_\mu(X)$

Remark: In the matrix case, this point is easy to illustrate. Consider a stochastic matrix $\mathcal{S}$ and a probability distribution ν on $\mathbb{R}^N$, with N equal to the column count of $\mathcal{S}$. Then for $1_{N\times 1} =: \mathbf{1}$, $\mathcal{S}\mathbf{1} = \mathbf{1}$ and $\nu'\mathbf{1} = 1$. Consider $\hat{\nu} = \mathcal{S}'\nu$. Then

$$\hat{\nu}'\mathbf{1} = (\mathcal{S}'\nu)'\mathbf{1} = \nu'\mathcal{S}\mathbf{1} = \nu'\mathbf{1} = 1$$

so the rows of $\mathcal{S}'$ take probability measures to probability measures.

Proof of part I. By the Perron-Frobenius theorem, $\mathcal{M}$ has largest eigenvalue $\lambda_0 = 1$ with corresponding left and right eigenvectors μ'_0 and $\iota = 1_{S\times 1}$, respectively.

Eigenvectors exhibit a scale symmetry $\mathcal{M}\iota = \iota \Leftrightarrow c\mathcal{M}\iota = c\iota$, $0 \neq c \in \mathbb{R}$, and equivalently $c^*\mu'_0 = c^*\mu'_0\mathcal{M}$ for any $0 \neq c^* \in \mathbb{R}$. Without loss of generality, we pick the the unitary scale normalization $c^* = c^{-1}$ so that $\iota'\mu_0 = 1$ so μ_0 is a probability distribution. Because the stochastic matrix restricts $c \equiv 1$, this choice of scale pins down the *absolute* scale $c^* = 1$.

Write the algebraic multiplicity of eigenvalue j as $\chi(\{j\})$, and the geometric multiplicity $g(\{j\})$. Denoting the largest eigenvalue $j = 0$, another appeal to the Perron-Frobenius theorem gives $\chi(\{0\}) = g(\{0\}) = 1$. We can now claim the following.

Lemma A.0.7 *$\mu_0\iota'$ is a rank-one projection.*

Proof of lemma 7.7 The linear operation $\mu_0\iota'$ is idempotent

$$\mu_0\iota'(\mu_0\iota') = \mu_0(\iota'\mu_0)\iota' = \mu_0\iota'$$

which immediately implies that $I - \mu_0\iota'$ is orthogonal to $\mu_0\iota'$,

$$(I - \mu_0\iota')\mu_0\iota' = \mu_0\iota' - \mu_0\iota' = \mathbf{0}$$

so $\mu_0\iota'$ and $I - \mu_0\iota'$ are projections. That $g(\{0\}) = 1$ implies that the subspace

$$E_0 := \{\hat{\mu} : (\mathcal{M}' - \lambda_0 I)\hat{\mu} = \mathbf{0}\}$$

has dimension one, hence $\text{rank}(\mu_0\iota') = 1$ and $\mu_0\iota'$ is a rank-one projection.

□

To prove part **I**. of proposition (7.1), note that if $\text{rank}(\mathcal{M}') = 1$, the result is trivial with $\mathscr{I}_0^{\perp}$ empty. Suppose $\text{rank}(\mathcal{M}') > 1$. Put

$$\mathcal{M}'_{\boldsymbol{\gamma}} = \mathcal{M}' - \mu_0\iota'$$

Denote $\text{Span}(\mu_0\iota') =: \mathscr{I}_0$. We intend to characterize $\mathscr{I}_0^{\perp} \subset \mathscr{R}(\mathcal{M}')$ in terms of $\mathcal{M}'_{\boldsymbol{\gamma}}$.

We can see that

$$\mathcal{M}'_{\boldsymbol{\gamma}}\mu_0\iota' = (\mathcal{M}' - \mu_0\iota')\mu_0\iota' = \mathcal{M}'(I_S - \mu_0\iota')\mu_0\iota' = \mathbf{0} \tag{6.a}$$

The second equality has used the right-eigenvector of the adjoint $\mathcal{M}'\mu_0 = \mu_0$, and the final equality follows given $\mu_0\iota'$ is a rank-one projection.

We verify that, in addition,

$$\mu_0\iota'\mathcal{M}'_{\boldsymbol{\gamma}} = (\mathcal{M}_{\boldsymbol{\gamma}}\iota\mu_0')' = ((\mathcal{M} - \iota\mu_0')\iota\mu_0')' = (\mathcal{M}(I_S - \iota\mu_0')\iota\mu_0')' = \mathbf{0} \tag{6.b}$$

using the right-eigenvector $\iota = \mathcal{M}\iota$, and the facts that $\text{rank}(\mu_0\iota') = \text{rank}(\iota\mu_0')$ and $\mu_0\iota'$ is a rank-one projection.

The image of $\mathcal{M}'$ takes the form $\mathscr{R}(\mathcal{M}') = \mathscr{I}_0 \bigoplus \text{Span}(\mathcal{M}'_{\boldsymbol{\gamma}})$, because in (6.a)-(6.b), we have shown $\mathscr{I}_0^{\perp} = \text{Span}(\mathcal{M}'_{\boldsymbol{\gamma}})$.

Now define $\mathcal{M}'_{\boldsymbol{\gamma},k} = (\mathcal{M}')^k - \mu_0\iota'$. By definition of the invariant measure μ_0, we have the right-eigenvector arguments for integer powers k,

$$\begin{aligned} (\mathcal{M}')^k\mu_0 &= \mu_0 \\ (\mathcal{M})^k\iota &= \iota \end{aligned} \tag{6.c}$$

Combining the first identity in (6.c) with the arguments (6.a), we obtain

$$\mathcal{M}'_{\boldsymbol{\gamma},k}\mu_0\iota' = [(\mathcal{M}')^k - \mu_0\iota']\mu_0\iota' = (\mathcal{M}')^k(I_S - \mu_0\iota')\mu_0\iota' = \mathbf{0}$$

Repeating the arguments in (6.b) using the second identity in (6.c), and the and the elementary fact $((\mathcal{M})^k)' = (\mathcal{M}')^k$,

$$\mu_0\iota'\mathcal{M}'_{\gamma,k} = ([(\mathcal{M}')^k - \mu_0\iota']'\iota\mu'_0)' = (\mathcal{M}^k(I_S - \iota\mu'_0)\iota\mu'_0)' = \mathbf{0}$$

Finally, for $0 < k \in \mathbb{N}$, it is clear that

$$\begin{aligned}(\mathcal{M}'_\gamma)^k &= (\mathcal{M}')^k(I_S - \mu_0\iota')^k \\ &= (\mathcal{M}')^k(I_S - \mu_0\iota') \\ &= (\mathcal{M}')^k - \mu_0\iota' \\ &= \mathcal{M}'_{\gamma,k}\end{aligned}$$

We conclude that for any integer $k \geq 1$, $(\mathscr{I}_0^\perp)_k = \text{Span}(\mathcal{M}'_{\gamma,k})$ with $\mathcal{M}'_{\gamma,k} = (\mathcal{M}'_\gamma)^k$ and $\mathcal{M}'_\gamma = \mathcal{M}' - \mu_0\iota'$. In particular, $\mathscr{R}((\mathcal{M}')^k) = \mathscr{I}_0 \bigoplus \text{Span}((\mathcal{M}'_\gamma)^k)$.

□

Before proving part II., we state some useful corollaries.

Corollary A.0.8 *$\mu_0\iota'$ and $\mathcal{M}'_\gamma$ commute.*

Corollary A.0.9 (Chapman-Kolmogorov) *The operators $\{(\mathcal{M}')^n, n \in \mathbb{Z}^+\}$ form an abelian semigroup under (matrix) multiplication. Every k- decomposition is contained in the semigroup. An identical statement is true for the operators $\{(\mathcal{M})^n, n \in \mathbb{Z}^+\}$.*

Proof Pick finite positive integers n_1, n_2 and denote $N = n_1 + n_2$. Then,

$$\begin{aligned}(\mathcal{M}')^{n_1}(\mathcal{M}')^{n_2} &= (\mu_0\iota' + \mathcal{M}'_\gamma)^N \\ &= \sum_{n=0}^{N} \binom{N}{n} (\mu_0\iota')^{N-n}(\mathcal{M}'_\gamma)^n = \mu_0\iota' + (\mathcal{M}'_\gamma)^N\end{aligned}$$

The first line is trivial, and shows the semigroup property $(\mathcal{M}')^{n_1}(\mathcal{M}')^{n_2} = (\mathcal{M}')^N$ and hence commutativity $(\mathcal{M}')^N = (\mathcal{M}')^{n_2}(\mathcal{M}')^{n_1}$. The second line applies the results from proposition (7.1) finitely many times to extend the semigroup property to the decomposition.

□

Corollary A.0.10 (γ- Semigroup) *The operators $\{(\mathcal{M}'_\gamma)^n, n \in \mathbb{Z}^+\}$ form an abelian semigroup under matrix multiplication.*

Corollary A.0.11 (Wold Time Series Representation) *Under the assumptions needed to decompose $(\mathcal{M}')^k = \mu_0\iota' + (\mathcal{M}'_\gamma)^k$, the classic Wold representation is justified.*

Proof First, recall each $\mathbb{R}^S \ni 1(x_{t,\cdot})$ is a degenerate probability distribution with mass on the coordinate of the realized element $x_{t,\cdot} \in X$. Recall the definitions $R_{n,t+1} = r_n \cdot X_{t+1}$, and hence $x_{t+1} = \iota'\mathcal{M}'1(x_{t,\cdot}) + u_{t+1}$. Equivalently,

$$1(x_t) = \mathcal{M}'1(x_{t-1}) + 1(u_t)$$

If $\mathcal{M}' \equiv \mu_0\iota'$, the decomposition is $r_n \cdot X_{t+1} = r_n \cdot \mu_0 + u_{t+1}$ for every t. Consider Rank$(\mathcal{M}') > 1$. Then,

$$\begin{aligned}
r_n \cdot X_{t+1} &= r_n \cdot (\mathcal{M}'1(x_t) + 1(u_{t+1})) \\
&= r_n \cdot \left([\mu_0\iota' + \mathcal{M}'_{\boldsymbol{\gamma}}]1(x_t) + 1(u_{t+1})\right) \\
&= r_n \cdot \mu_0 + r_n \cdot \left(\mathcal{M}'_{\boldsymbol{\gamma}}1(x_t) + 1(u_{t+1})\right) \\
&= r_n \cdot \mu_0 + r_n \cdot \left(\mathcal{M}'_{\boldsymbol{\gamma}}\left(\mathcal{M}'1(x_{t-1}) + 1(u_t)\right) + 1(u_{t+1})\right) \\
&= r_n \cdot \mu_0 + r_n \cdot \left(\mathcal{M}'_{\boldsymbol{\gamma}}([\mu_0\iota' + \mathcal{M}'_{\boldsymbol{\gamma}}]1(x_{t-1}) + 1(u_t)) + 1(u_{t+1})\right) \\
&= r_n \cdot \mu_0 + r_n \cdot \left(\mathcal{M}'_{\boldsymbol{\gamma}}(\mathcal{M}'_{\boldsymbol{\gamma}}1(x_{t-1}) + 1(u_t)) + 1(u_{t+1})\right) \\
&= r_n \cdot \mu_0 + r_n \cdot \left(\mathcal{M}'_{\boldsymbol{\gamma}}(\mathcal{M}'_{\boldsymbol{\gamma}}\left(\mathcal{M}'1(x_{t-2}) + 1(u_{t-1})\right) + 1(u_t)) + 1(u_{t+1})\right) \\
&\vdots \\
&= r_n \cdot \mu_0 + r_n \cdot \sum_{s=0}^{\infty} (\mathcal{M}'_{\boldsymbol{\gamma}})^s 1(u_{t+1-s})
\end{aligned}$$

□

Remark: The Wold representation expresses the path $\{z\}(\omega)$ in terms of its martingale-difference (i.e., white-noise) basis. A finite sample version is obtained as a special case. For generalized Wold representations of isometric operators, see Severino (2014).

Remark: The $\mu_0\iota'$ shocks are *permanent* in the sense that $\mu_0\iota'1(u_t) = (\mathcal{M}')^N\mu_0\iota'1(u_t)$ for every horizon N.

Corollary A.0.12 (Preservation of Probabilities) *The rows of $\mathcal{M}_{\boldsymbol{\gamma}}$ sum to zero. When Rank$(\mathcal{M}) > 1$, there are nontrivial payoff vectors $d_n \in F(K)$ such that*

$$|e_i \cdot \mathcal{M}'_{\boldsymbol{\gamma}} d_n| > 0$$

for some $i \in \{1, ..., S\}$

Proof Put $1 = 1_{S\times 1}\,(= \iota)$. Recall the right eigenvector of $\mathcal{M}\iota = \iota$, so that

$$1 = \mathcal{M}\iota = (\iota\mu'_0 + \mathcal{M}_{\boldsymbol{\gamma}})\iota = \iota\mu'_0\iota + \mathcal{M}_{\boldsymbol{\gamma}}\iota = 1 + \mathcal{M}_{\boldsymbol{\gamma}}\iota$$

Clearly, $\mathcal{M}_{\boldsymbol{\gamma}}\iota = \mathbf{0}$.

Given the invariant $\mu'_0 = \mu'_0\mathcal{M}$, the matrix $\mathcal{M}_{\boldsymbol{\gamma}}$ is identically zero when $\mathcal{M}$ is identically $(\mu_0\iota')'$. Given μ_0, a necessary and sufficient condition for Rank$(\mathcal{M}) = 1$ is that the rows of $\mathcal{M}$ are constant multiples of μ'_0.

Take $\mathcal{M}'$ such that Rank$(\mathcal{M}') > 1$, and consider d_n s.t. $|d_n \cdot e_i| > 0$ for some $i \in \{1, ..., S\}$. Then $\mathcal{M}'d_n = \mu_0\iota'd_n + \mathcal{M}'_{\boldsymbol{\gamma}}d_n$ and $0 \leq |e_i \cdot (\mathcal{M}' - \mu_0\iota')d_n| = |e_i \cdot \mathcal{M}'_{\boldsymbol{\gamma}}d_n|$ for every $1 \leq i \leq S$.

Suppose for a contradiction that $e_i \cdot (\mathcal{M}' - \mu_0\iota')d_n = 0$ for every i. In other words, $\mathcal{M}'d_n = \mu_0\iota'd_n$ for every nonzero $d_n \in F_d(K)$. Because $\mathcal{M}'\mu_0 = \mu_0$, these together imply

$$\begin{aligned} \mathcal{M}'(\mu_0 - d_n) &= \mu_0(1 - \iota' d_n) \\ &= \mu_0\iota'(\tfrac{1}{S}\iota - d_n) \end{aligned} \tag{1.a}$$

The condition $\text{Rank}(\mathcal{M}') > 1$ ensures $\mathcal{M}'$ is not identically $\mu_0\iota'$. In particular, in $\mathscr{R}(\mathcal{M}') = \mathscr{I}_0 \bigoplus \mathscr{I}_0^{\perp}$, the $\mathscr{I}_0^{\perp}$ is nonempty. So, (1.a) implies $d_n = \mu_0 = \frac{1}{S}\iota$, the uniform distribution over Z_S. If μ_0 is not uniform, we are done. If μ_0 is uniform, then $\mathcal{M}'\iota = \iota$, so $\mathcal{M}'$ is doubly stochastic. As a result, $\mathcal{M}'d_n = \mu_0\iota'd_n = \frac{1}{S}\iota\iota'd_n$, which implies each row of $\mathcal{M}'$ is identically the uniform distribution. Hence, contrary to our assumption, $\text{Rank}(\mathcal{M}') = 1$. Thus, for some $i \in \{1, ..., S\}$, $0 < |e_i \cdot (\mathcal{M}' - \mu_0\iota')d_n| = |e_i \cdot \mathcal{M}'_{\boldsymbol{\gamma}} d_n|$.

□

Remark: $\mathcal{M}'_{\boldsymbol{\gamma}}$ redistributes probability within the transition dynamics. Relative to dynamics under $\mu_0\iota'$, $\mathcal{M}'_{\boldsymbol{\gamma}}$ generates non-negligible distortions to asset payoffs locally in time.

It turns out these distortions are limited to risky assets.

Corollary A.0.13 (State Dependent Payoffs) *Every nontrivial constant payoff is in the image space of $\mu_0\iota'$. In particular, the uniform distribution is never the range of $\mathcal{M}'_{\boldsymbol{\gamma}}$.*

Proof Take $\text{Rank}(\mathcal{M}') > 1$ so $\text{Rank}(\mathcal{M}'_{\boldsymbol{\gamma}}) \geq 1$. Put

$$Ker(\mathcal{M}_{\boldsymbol{\gamma}}) := \{v \in C_K(X) : \mathcal{M}_{\boldsymbol{\gamma}} v = 0_{S\times 1}\}$$

Recall $1_{S\times 1} = \iota$ and $c\mathcal{M}_{\boldsymbol{\gamma}}\iota = 0$ for any scalar $c \in \mathbb{R}$. It follows that

$$b_0 := \{v \in C_K(X) : v = c\,\iota\,,\ 0 \neq c \in \mathbb{R}\} \subset Ker(\mathcal{M}_{\boldsymbol{\gamma}})$$

Now, using[1]

$$\mathscr{R}(\mathcal{M}'_{\boldsymbol{\gamma}}) = Ker(\mathcal{M}_{\boldsymbol{\gamma}})^{\perp}$$

we conclude that $b_0 \nsubseteq \mathscr{R}(\mathcal{M}'_{\boldsymbol{\gamma}})$. In particular, $b_0 \subset \mathscr{I}_0$.

□

A.0.3 Convergence Rates

To study convergence, we review a few definitions and known results. A handful of purely analytical definitions are relegated to section (7.6.1.).

Definition: The *total variation* of a signed measure is

$$\| \eta \|_{TV} := \sup_{A\in\mathcal{B}(X)} |\eta(A)|$$

[1] e.g., Luendberger, sec 6.6 pp 155, theorem 1.

Definition: The total variation *distance* between two probability measures μ, ν is

$$d(\mu,\nu)_{TV} = \| \mu - \nu \|_{TV} = \sup_{b \in (2^X)^{\mathbb{N}}} |\mu(b) - \nu(b)|$$

which itself takes values in $[0,1]$.

Lemma A.0.14 *The set of probability measures F_μ is metrized by total variation $d = d_{TV}$. In particular (F_μ, d_{TV}) inherits the topology induced by total variation. Now write $\sigma(F_\mu)$ for the Borel σ -field on F_μ with open sets generated by the total variation distance d_{TV}. Then $(F_\mu, \sigma(F_\mu))$ is a measurable space.*

Proof Den Hollander (2011), p.13 II.1

□

Lemma A.0.15 *On countable state space, total variation distance is equivalent to*

$$\sup_{b \in (2^X)^{\mathbb{N}}} |\mu(b) - \nu(b)| = \frac{1}{2} \sum_{a \in X} |\mu(a) - \nu(a)|$$

Proof Levin, Peres and Wilmer, 2008, Proposition 4.2.

□

An immediate corollary of lemma (7.7) is that for finite X, total variation is an ℓ^1 *norm*

$$\begin{aligned} \sum_{a \in 2^X} |h_0(a) - \mu_0(a)| &= \sum_{a \in 2^X} |\mathsf{h}_0(a) - 1| \mu_0(a) \\ &= \| \mathsf{h}_0 - \iota \|_{\ell^1(\mu_0)} \end{aligned}$$

under μ_0, where the elementwise quotient $\mathsf{h}_0 = h_0/\mu_0$ is the likelihood of h_0 with respect to μ_0.

Definition: The operator norm $\| \cdot \|_{Op}$ for operator $\mathbb{T}$ is

$$\| \mathbb{T} \|_{Op} := \sup_{\{v \in \mathcal{D}(\mathbb{T}),\, \|v\|_{\ell^1} = 1\}} \| \mathbb{T} v \|_{\ell^1}$$

Lemma A.0.16 *For continuous linear operators $\mathbb{T}, \mathbb{S}$*

$$\| \mathbb{T}'\mathbb{T} \|_{Op} = \| \mathbb{T} \|_{Op}^2 \quad \text{(a.)}$$

$$\| \mathbb{S}\mathbb{T} \|_{Op} \leq \| \mathbb{S} \|_{Op} \| \mathbb{T} \|_{Op} \quad \text{(b.)}$$

Proof
(a.) Lax (2002), Theorem 14., section 19.7, p. 222
(b.) Lax (2002) Theorem 8., section 15.4, p. 168

□

Proof of Prop. 7.1 part II. We need only consider the nontrivial case $\text{Rank}(\mathcal{M}') > 1$. Pick an initial probability over X given by $\hat{h}_0$. Then $\hat{h}'_{0,k} = \hat{h}'_0 \mathcal{M}^k$, equivalently $(\mathcal{M}')^k \hat{h}_0 = \hat{h}_{0,k}$. Define the likelihood $\mathsf{h}_0 = \hat{h}/\mu_0$ with the quotient elementwise as before. Consider first the case $\hat{h}_0 \neq \mu_0$. We are interested in $\| (\mathcal{M}')^k \hat{h}_0 - \mu_0 \|_{TV}$. From part I.,

$$\begin{aligned} (\mathcal{M}')^k \hat{h}_0 - \mu_0 &= \mu_0 \iota' \hat{h}_0 + (\mathcal{M}'_\gamma)^k \hat{h}_0 - \mu_0 \\ &= (\mathcal{M}'_\gamma)^k \hat{h}_0 + \mu_0(\iota' \hat{h}_0 - 1) \\ &= (\mathcal{M}'_\gamma)^k \hat{h}_0 \end{aligned}$$

The third equality is simply $\iota'\hat{h}_0 = 1$ because $\hat{h}_0$ is a probability measure. We can now write

$$\begin{aligned} 2 \parallel (\mathcal{M}')^k \hat{h}_0 - \mu_0 \parallel_{TV}^2 &= 2 \parallel (\mathcal{M}'_{\boldsymbol{\gamma}})^k \hat{h}_0 \parallel_{TV}^2 \\ &= \parallel (\mathcal{M}'_{\boldsymbol{\gamma}})^k \mathsf{h}_0 \parallel_{\ell^1(\mu_0)}^2 \\ &\leq \parallel (\mathcal{M}'_{\boldsymbol{\gamma}})^{2k} \mathbf{1} \parallel_{\ell^1(\mu_0)} \parallel \mathsf{h}_0^2 \parallel_{\ell^1(\mu_0)} \end{aligned}$$

using Cauchy-Schwarz.

We showed in corollary 7.13 that the condition $\mathcal{M}_{\boldsymbol{\gamma}}\iota = \mathcal{M}'_{\boldsymbol{\gamma}}\iota$ imposes that $\mathcal{M}$ is doubly stochastic and therefore that μ_0 is uniform $\mu_0 = \frac{1}{S}\iota$ and have ruled this case out. Because $\mathbf{1}(c) := \{v : v = c\mathbf{1}\, 0 \neq c \in \mathbb{R}\} \subset \mathrm{Ker}(\mathcal{M}_{\boldsymbol{\gamma}})$, we conclude $\mathcal{M}'_{\boldsymbol{\gamma}}\mathbf{1}$ is not trivial. Proceeding

$$\begin{aligned} \parallel (\mathcal{M}'_{\boldsymbol{\gamma}})^{2k} \mathbf{1} \parallel_{\ell^1(\mu_0)} \parallel \mathsf{h}_0^2 \parallel_{\ell^1(\mu_0)} &= 2 \parallel (\mathcal{M}'_{\boldsymbol{\gamma}})^{2k} \mathbf{1} \parallel_{\ell^1(\mu_0)} \parallel \hat{h}_0^2 \parallel_{TV} \\ &\leq 2 \parallel (\mathcal{M}'_{\boldsymbol{\gamma}})^{2k} \mathbf{1} \parallel_{\ell^1(\mu_0)} \end{aligned} \tag{6.1.a}$$

because the total variation distance is no greater than one.

The operator $\mathcal{M}'_{\boldsymbol{\gamma}}$ is a contraction for every $k \in \mathbb{N}$ and is therefore uniformly bounded. Because it is also linear, by the open mapping theorem it is continuous in the operator norm topology (e.g., Lax (2002), Theorem 12. p. 170). Now define $\mathbf{M}_\gamma := \mathcal{M}_{\boldsymbol{\gamma}}\mathcal{M}'_{\boldsymbol{\gamma}}$. The operator $\mathbf{M}_\gamma$ is linear, uniformly bounded in k and symmetric. We have

$$\begin{aligned} 2 \parallel (\mathcal{M}'_{\boldsymbol{\gamma}})^{2k} \mathbf{1} \parallel_{\ell^1(\mu_0)} &\leq 2 \max_{||v||=1} \parallel (\mathcal{M}'_{\boldsymbol{\gamma}})^{2k} v \parallel_{\ell^1(\mu_0)} \\ &= 2 \parallel (\mathcal{M}'_{\boldsymbol{\gamma}})^{2k} \parallel_{Op} \\ &\leq 2 \parallel \mathcal{M}'_{\boldsymbol{\gamma}} \parallel_{Op}^{2k} \\ &= 2 \parallel \mathcal{M}_{\boldsymbol{\gamma}}\mathcal{M}'_{\boldsymbol{\gamma}} \parallel_{Op}^{k} \\ &= 2 \parallel \mathbf{M}_\gamma \parallel_{Op}^{k} \\ &\leq 2\rho(\mathbf{M}_\gamma)^k \end{aligned} \tag{6.1.b}$$

where $\rho(\mathbf{M}_\gamma)$ is the spectral radius of $\mathbf{M}_\gamma$, and we have used submultiplicativity and part (a.) in lemma 7.18.

When $\mathrm{rank}(\mathcal{M}') > 1$, the symmetric positive definite operator $\mathcal{M}\mathcal{M}'$ has at least two eigenvalues. In particular, $\rho(\mathbf{M}_\gamma)$ is bounded above by the second largest eigenvalue of $\mathcal{M}\mathcal{M}'$. Because $\mathcal{M}\iota = \iota$, the largest eigenvalue of $\mathcal{M}\mathcal{M}'$ is one. The second largest eigenvalue ζ of $\mathbf{M}$ is strictly less than one. Moreover, if $\mathcal{M}\psi = \zeta^{1/2}\psi$ for $\mathfrak{Re}(\zeta^{1/2}) < 1$, then

$$\mathcal{M}\psi = \iota\mu_0'\psi + \mathcal{M}_{\boldsymbol{\gamma}}\psi = \mathcal{M}_{\boldsymbol{\gamma}}\psi = \zeta^{1/2}\psi$$

Consider $\varphi'\mathcal{M} = \varphi'\iota\mu_0' + \varphi'\mathcal{M}_{\boldsymbol{\gamma}}$ and $\varphi'\psi = \zeta^{1/2}$, so that

$$\parallel \mathcal{M}'\varphi \parallel = \parallel \mu_0\iota'\varphi + \mathcal{M}'_{\boldsymbol{\gamma}}\varphi \parallel = \parallel \mathcal{M}'_{\boldsymbol{\gamma}}\varphi \parallel \leq |\zeta^{1/2}||\varphi'\psi| = \zeta$$

We conclude

$$\rho(\mathbf{M}_\gamma) \leq \zeta < 1 \tag{6.1.c}$$

Combining (6.1.a) - (6.1.c), we have

$$\| (\mathcal{M}')^k \hat{h}_0 - \mu_0 \|_{TV} \leq \zeta^{k/2} \longrightarrow 0 \quad k \to \infty$$

Now, the case not yet considered is when $h_0 = \mu_0$ but $\mathcal{M}$ is not identically $\iota\mu_0'$. We are interested in the rate at which

$$\| \iota' \mathcal{M}^k f - \mu_0' f \|_{TV} \longrightarrow 0 \qquad k \to \infty$$

However, it is straightforward to bound this deviation using the same radius $\rho(\mathbf{M}_\gamma)$. Expediting the argument using details in (6.1.a)-(6.1.c),

$$\begin{aligned} 2 \| \iota' \mathcal{M}^k f - \mu_0' f \|_{TV}^2 &= 2 \| (\iota' \mathcal{M}^k - \mu_0') f \|_{TV}^2 \\ &= \| (\iota' \mathcal{M}^k - \mu_0') f \|_{\ell^1}^2 \\ &\leq \| \mathcal{M}_{\boldsymbol{\gamma}}^{2k} \|_{\ell^1} \| f^2 \|_{\ell^1} \\ &\leq 2 \| \mathcal{M}_{\boldsymbol{\gamma}} \|_{Op}^{2k} \\ &= 2 \| \mathbf{M}_\gamma \|_{Op}^k \\ &\leq 2\rho(\mathbf{M}_\gamma)^k \end{aligned}$$

Hence,

$$\| \iota' \mathcal{M}^k f - \mu_0' f \|_{TV} \leq \zeta^{k/2} \longrightarrow 0 \quad k \to \infty$$

□

Remark: A square root is natural when ζ is viewed as a singular value of $\mathcal{M}'$, in which case the corresponding eigenvalue of $\mathcal{M}'$ is $\zeta^{1/2}$.

Remark: The generator $\mathcal{L}$ inherits a decomposition,

$$(\mathcal{L}'h)(x) = (\mathcal{M}_{\boldsymbol{\gamma}}' h)(x) + ((\mu_0 \iota' - I)h)(x)$$

from proposition 7.1. A function g is γ - *harmonic* when

$$\begin{aligned} 0 &= ((\mathcal{M}_{\boldsymbol{\gamma}}' - I)g)(x) \\ &= (\mathcal{L}'g)(x) - (\mu_0 \iota' g)(x) \end{aligned}$$

emphasizing that such a g has conditionally mean-zero transitory contributions.

Definition: ν - harmonics A function h such that $(\mathcal{L}h)(x) = 0$ for every x is *harmonic* with respect the measure $\nu_{\mathcal{M}}$ on X' when in addition, $\langle \mathcal{L}h, 1 \rangle_{\nu(\mathcal{M})} = \langle h, \mathcal{L}^* \rangle_{\nu(\mathcal{M})}$.

Proposition A.0.17 *Then,*

- *Every harmonic function h with respect to $\nu_{\mathcal{M}}$ is a martingale under $\mathbb{P}_{\nu_{\mathcal{M}}}$ as a function of the initial conditions*
- *Every martingale is contained in the kernel of $\mathcal{M}'_{\gamma}$ - i.e., every $\nu_{\mathcal{M}}$ - harmonic function is γ -harmonic when $\mathcal{M}_{\gamma}$ decomposes $\mathcal{M}$ and $\nu'_{\mathcal{M}} = \nu'_{\mathcal{M}}\mathcal{M}$*

Proof Part *i*) is given by Doob (1959). Part *ii*) follows from lemma (7.8) and $\text{Ker}(\mathcal{A}) = \mathscr{R}(\mathcal{A}')^{\perp}$.

□

A.0.4 Covariance Matrix of Returns

Define

$$\mathscr{E}(\phi) = \| \phi \|^2_{\ell^2(\mu_0)} - \| \mathcal{M}'\phi \|^2_{\ell^2(\mu_0)}$$

We will make use of a variational representation of the spectral gap given by Diaconis and Strook 1991,

$$\zeta := 1 - \lambda_1 = \inf_{\phi} \left\{ \frac{\mathscr{E}(\phi)}{\| \phi \|^2_{\ell^2(\mu_0)}} \text{ s.t. } \| \phi \|^2_{\ell^2(\mu_0)} > 0 \right\}$$

We now unpack the contents of $\mathscr{E}(\phi)$ in terms of realized returns data.

Inner Product To compare any two sequences a and b in Z, we consider the natural inner product $\langle \{z_a\}, \{z_b\} \rangle$ on ℓ^2 and the normalized inner product $\langle \{z_a\}, \{z_b\} \rangle_{\mu_0}$.

Lemma A.0.18 *The path space appended with either of these inner products is a Hilbert space.*

Proof Strook, 2014, p. 139

□

We occasionally write ϕ for paths of the Markov chain. Recall $\mathscr{E}_M(\phi) = \mathscr{E}_M(\phi, \phi) := \langle \phi, (I - \mathcal{M}\mathcal{M}')\phi \rangle_{\mu}$, and that for real operators B, the adjoint operator is B', $\langle B\phi, \phi \rangle_{\mu} = \langle \phi, B'\phi \rangle_{\mu}$. $\mathscr{E}_M(\phi)$ rearranges,

$$\begin{aligned} \mathscr{E}_M(\phi, \phi) &= \langle \phi, (I - \mathcal{M}\mathcal{M}')\phi \rangle_{\mu} \\ &= \langle \phi, \phi \rangle_{\mu} - \langle \phi, \mathcal{M}\mathcal{M}'\phi \rangle_{\mu} \\ &= \mathbb{E}[\phi\phi'] - \langle \mathcal{M}'\phi, \mathcal{M}'\phi \rangle_{\mu} \\ &= \mathbb{E}[\phi\phi'] - \mathbb{V}(\mathcal{M}'\phi) \end{aligned}$$

Furthermore, $\mathbb{E}[\phi\phi'] = \langle \phi, \phi \rangle_{\mu} = \mathbb{V}(\phi) + \mu_0\mu'_0$, where $\mathbb{V}(\phi)$ is the variance of ϕ, and $\mathbb{V}(\mathcal{M}'\phi)$ is the variance of the conditional mean of ϕ.

Define the lag operator $L\phi_{t+1} = \phi_t$. Each path realization is an ℓ^2 sequence of the form $\phi = \{z\}(\omega) = (a_0 u_0), (a_1 u_1), (a_2 u_2), ... =: A_0 W_0$. The u_j are serially uncorrelated $\langle u_j, u_k \rangle = 0$ for $j \neq k$. The lag operator L maps $Z \to Z$ via the shift $LA_0W_0 \mapsto A_0W_{-1}$. The adjoint of the lag operator L^* maps $L^*A_0W_0 \mapsto A_1W_0$. Heuristically,

$$\begin{aligned}\langle LA_tW_t, A_tW_t \rangle &= \langle A_tW_{t-1}, A_tW_t \rangle = A'_{t+1}A_t \\ \langle A_tW_t, L^*A_tW_t \rangle &= \langle A_tW_t, A_{t+1}W_t \rangle = A'_tA_{t+1}\end{aligned}$$

as required.

Lemma A.0.19 (Lag Operator Isometry) *With this construction, $L : Z \to Z$ is an isometry on $(Z, \mu_0, ||\cdot||_{\ell^2})$.*

Proof We sketch a proof here for intuition. A detailed proof and discussion is given in Severino (2014). An isometric operator leaves norms and inner products unchanged when applied symmetrically. Notice

$$\begin{aligned}\langle LA_tW_t, LA_tW_t \rangle &= \langle A_tW_{t-1}, A_tW_{t-1} \rangle \\ &= A'_tA_t \\ &= \langle A_tW_t, A_tW_t \rangle\end{aligned}$$

The same argument shifting coefficients A_t in place of W_t shows the adjoint L^* is also an isometry.

□

Proposition A.0.20 *The form $\mathscr{E}_M(\phi)$ measures the unconditional variance of the forecast errors generated by $\mathcal{M}'$.*

Proof Recall

$$u_{t+1} = \phi_{t+1} - \mathcal{M}'\phi_t = (I - \mathcal{M}'L)\phi_{t+1}$$

The forecast errors are conditionally mean-zero by construction. The unconditional mean is also zero,

$$\begin{aligned}\mathbb{E}_0[u_{t+1}] &= \mu_0\iota'\phi_{t+1} - \mu_0\iota'\mathcal{M}'L\phi_{t+1} \\ &= \mu_0\iota'\phi_{t+1} - \mu_0\iota'L\phi_{t+1} - \mu_0\iota'\mathcal{M}'_\gamma L\phi_{t+1} \\ &= \mu_0\iota'(\phi_{t+1} - \phi_t) \\ &= 0 \qquad (7.10.a)\end{aligned}$$

because the span of $\mathcal{M}'_\gamma$ is orthogonal to $\mathscr{I}_0$ and the lag operator is identity on $\mathscr{I}_0$.

Using (7.10.a), the unconditional variance is of the form

$$\begin{aligned}
\mathbb{V}(u) = \langle u, u\rangle_\mu &= \langle (I - \mathcal{M}'L)\phi, (I - \mathcal{M}'L)\phi\rangle_\mu \\
&= \langle \phi, (I - \mathcal{M}'L)'(I - \mathcal{M}'L)\phi\rangle_\mu \\
&= \langle \phi, (I - I\mathcal{M}'L - L'\mathcal{M}I + L'\mathcal{M}\mathcal{M}'L)\phi\rangle_\mu \\
&= \mathbb{E}[\phi\phi'] + \langle \mathcal{M}'L\phi, \mathcal{M}'L\phi\rangle_\mu - \langle \phi, \mathcal{M}'L\phi\rangle_\mu - \langle \phi, L'\mathcal{M}\phi\rangle_\mu \\
&= \mathbb{E}[\phi\phi'] + \mathbb{V}(\mathcal{M}'\phi) - \langle \phi, \mathcal{M}'L\phi\rangle_\mu - \langle \phi, L'\mathcal{M}\phi\rangle_\mu \\
&= \mathbb{E}[\phi\phi'] + \mathbb{V}(\mathcal{M}'\phi) - \langle \phi, \mathcal{M}'L\phi\rangle_\mu - \langle \mathcal{M}'L\phi, \phi\rangle_\mu
\end{aligned}$$

where we have used the isometry of the lag operator L to get from the fourth to the fifth line, and the adjoint $(\mathcal{M}'L)' = L'\mathcal{M}$. Now

$$\begin{aligned}
\langle \mathcal{M}'L\phi, \phi\rangle_\mu &= \langle \mathcal{M}'\phi_t, \phi_{t+1}\rangle_\mu \\
&= \langle \phi_{t+1} - u_{t+1}, \phi_{t+1}\rangle_\mu \\
&= \langle \phi_{t+1}, \phi_{t+1}\rangle_\mu - \langle u_{t+1}, \phi_{t+1}\rangle_\mu
\end{aligned}$$

Using the Wold representation for the second term in the last equality,

$$\begin{aligned}
\langle u_{t+1}, \phi_{t+1}\rangle_\mu &= \langle u_{t+1}, \mu_0\iota' + \sum_{s=0}^{\infty} (\mathcal{M}_\gamma)^s\, u_{t+1-s}\rangle_\mu \\
&= \langle u_{t+1}, u_{t+1}\rangle_\mu
\end{aligned}$$

so

$$\langle \mathcal{M}'\phi_t, \phi_{t+1}\rangle_\mu = \langle \phi_{t+1}, \phi_{t+1}\rangle_\mu - \langle u_{t+1}, u_{t+1}\rangle_\mu$$

Applying a symmetric argument to $\langle \phi, \mathcal{M}'L\phi\rangle_\mu$, we have

$$\langle \phi_t, \mathcal{M}'\phi_{t+1}\rangle_\mu = \langle \phi_{t+1}, \phi_{t+1}\rangle_\mu - \langle u_{t+1}, u_{t+1}\rangle_\mu$$

Consolidating terms gives

$$\begin{aligned}
\langle u, u\rangle_\mu &= \mathbb{E}[\phi\phi'] + \mathbb{V}(\mathcal{M}'\phi) - \langle \phi, \mathcal{M}'\phi\rangle_\mu - \langle \mathcal{M}'\phi, \phi\rangle_\mu \\
&= \mathbb{E}[\phi\phi'] + \mathbb{V}(\mathcal{M}'\phi) - 2\langle \phi, \phi\rangle_\mu + 2\langle u, u\rangle_\mu \\
&\Longrightarrow \\
-\langle u, u\rangle_\mu &= \mathbb{V}(\mathcal{M}'\phi) - \mathbb{E}[\phi\phi'] \\
&= -\mathscr{E}_M(\phi, \phi)
\end{aligned}$$

□

Remark: Explicit time indices indicate material distinctions. For example, we have not shown $\langle \mathcal{M}'\phi, \phi\rangle = \langle \phi, \mathcal{M}'\phi\rangle$, rather $\langle \mathcal{M}'\phi_t, \phi_{t+1}\rangle = \langle \phi_{t+1}, \mathcal{M}'\phi_t\rangle$.

Wold Representations for Second Moments

For $\epsilon > 0$, each $\varsigma \in (0, 1+\epsilon]$ defines an operator

$$(\mathcal{R}_{\boldsymbol{\gamma}}(\varsigma)\phi)(x) = \varsigma^{-1} \sum_{n=0}^{\infty} \left(\varsigma^{-1}(\mathbf{M}_{\gamma}\phi)(x)\right)^n$$

We will also consider the *resolvent* as a *function* of ς,

$$\mathcal{R}_{\boldsymbol{\gamma}}(\varsigma) = \varsigma^{-1} \sum_{n=0}^{\infty} \left(\varsigma^{-1}\mathbf{M}_{\gamma}\right)^n$$

Because $\rho(\mathbf{M}_{\gamma}) < 1$, we have $\mathcal{R}_{\boldsymbol{\gamma}}(\varsigma) = (\varsigma I - \mathbf{M}_{\gamma})^{-1}$ for $\varsigma \in (\rho(\mathbf{M}_{\gamma}), 1+\epsilon]$, and

$$\mathcal{R}_{\boldsymbol{\gamma}}(1)(I - \mathbf{M}_{\gamma}) = I$$

A proof for the case of Neumann series is given in Lax ((2010), Theorem 3 p. 195).

A.0.5 Identification of Spectral Gap from Realized Returns

Proposition A.0.21 (Identification) *Consider the Markov environment above with rank*$(\mathcal{M}) > 1$. *The singular value decomposition of a panel of realized returns can be expressed in terms of the Markov transition*

$$\Lambda_{PCA} \doteqdot \mathsf{V}' D_{1-\lambda} \Sigma \mathsf{V}$$

where $\doteqdot$ *reads "unitarily equivalent" and where*

$$\mathcal{R}_{\gamma}(1) = \mathsf{U} D_{1-\lambda} \mathsf{U}'$$
$$(D_{1-\lambda})_{i,j} = \begin{cases} \frac{1}{1-\lambda_j} & i = j \\ 0 & i \neq j \end{cases}$$

Proof Recall that for any integer k, $\mu_0 \iota' \mathcal{M}_{\boldsymbol{\gamma}}^k = \mathbf{0}$, and for any $s > 0$, $\langle u_t, u_{t-s} \rangle = 0$. In particular, we have

$$\begin{aligned}
\mathbb{V}(R_t) &= \left\langle R_t - \overline{R}, R_t - \overline{R} \right\rangle \\
&= \langle \sum_{s=0}^{\infty} (\mathcal{M}_{\boldsymbol{\gamma}}')^s u_{t-s}, \sum_{s=0}^{\infty} (\mathcal{M}_{\boldsymbol{\gamma}}')^s u_{t-s} \rangle \\
&= \langle u_t, u_t \rangle + \langle \sum_{s=1}^{\infty} (\mathcal{M}_{\boldsymbol{\gamma}}')^s u_{t-s}, \sum_{s=1}^{\infty} (\mathcal{M}_{\boldsymbol{\gamma}}')^s u_{t-s} \rangle \\
&= \langle u_t, u_t \rangle + \left\langle \mathcal{M}_{\boldsymbol{\gamma}}' u_{t-1}, \mathcal{M}_{\boldsymbol{\gamma}}' u_{t-1} \right\rangle + \langle \sum_{s=2}^{\infty} (\mathcal{M}_{\boldsymbol{\gamma}}')^s u_{t-s}, \sum_{s=2}^{\infty} (\mathcal{M}_{\boldsymbol{\gamma}}')^s u_{t-s} \rangle
\end{aligned}$$

Now,

$$\begin{aligned}\left\langle \mathcal{M}'_{\boldsymbol{\gamma}} u_{t-1}, \mathcal{M}'_{\boldsymbol{\gamma}} u_{t-1} \right\rangle &= \left\langle \mathcal{M}'_{\boldsymbol{\gamma}} L u_t, \mathcal{M}'_{\boldsymbol{\gamma}} L u_t \right\rangle \\ &= \left\langle \mathcal{M}'_{\boldsymbol{\gamma}} u_t, \mathcal{M}'_{\boldsymbol{\gamma}} u_t \right\rangle \\ &= \left\langle \mathcal{M}_{\boldsymbol{\gamma}} \mathcal{M}'_{\boldsymbol{\gamma}} u_t, u_t \right\rangle \\ &= \langle \mathbf{M}_\gamma u_t, u_t \rangle\end{aligned}$$

using the lag operator isometry and the adjoint operation. Then,

$$\mathbb{V}(R_t) = \langle u_t, u_t \rangle + \langle \mathbf{M}_\gamma u_t, u_t \rangle + \left\langle (\mathcal{M}'_{\boldsymbol{\gamma}})^2 u_{t-2}, (\mathcal{M}'_{\boldsymbol{\gamma}})^2 u_{t-2} \right\rangle + ...$$

Similarly for the second order case,

$$\begin{aligned}\left\langle (\mathcal{M}'_{\boldsymbol{\gamma}})^2 u_{t-2}, (\mathcal{M}'_{\boldsymbol{\gamma}})^2 u_{t-2} \right\rangle &= \left\langle (\mathcal{M}'_{\boldsymbol{\gamma}})^2 L L u_t, (\mathcal{M}'_{\boldsymbol{\gamma}})^2 L L u_t \right\rangle \\ &= \left\langle (\mathcal{M}'_{\boldsymbol{\gamma}})^2 L u_t, (\mathcal{M}'_{\boldsymbol{\gamma}})^2 L u_t \right\rangle \\ &= \left\langle (\mathcal{M}'_{\boldsymbol{\gamma}})^2 u_t, (\mathcal{M}'_{\boldsymbol{\gamma}})^2 u_t \right\rangle \\ &= \left\langle \mathcal{M}_{\boldsymbol{\gamma}} (\mathcal{M}'_{\boldsymbol{\gamma}})^2 u_t, \mathcal{M}'_{\boldsymbol{\gamma}} u_t \right\rangle \\ &= \left\langle (\mathcal{M}_{\boldsymbol{\gamma}})^2 (\mathcal{M}'_{\boldsymbol{\gamma}})^2 u_t, u_t \right\rangle \\ &= \left\langle (\mathbf{M}_\gamma)^2 u_t, u_t \right\rangle\end{aligned}$$

using associativity. Continuing,

$$\begin{aligned}\mathbb{V}(R_t) &= \langle u_t, u_t \rangle + \langle \mathbf{M}_\gamma u_t, u_t \rangle + \left\langle (\mathbf{M}_\gamma)^2 u_t, u_t \right\rangle + \left\langle (\mathcal{M}'_{\boldsymbol{\gamma}})^3 u_{t-3}, (\mathcal{M}'_{\boldsymbol{\gamma}})^3 u_{t-3} \right\rangle + ... \\ &\vdots \\ &= \lim_{N \to \infty} \sum_{s=0}^{N} \langle (\mathbf{M}_\gamma)^s u_t, u_t \rangle\end{aligned}$$

This sum is absolutely convergent because the eigenvalues are bounded inside the unit circle uniformly in parameter $k \in \mathbb{N}$.

Write $\sum_{n=0}^{\infty} m_n = \int_{\mathbb{R}} m(n) \nu(dn)$ for the counting measure $\hat{\nu}(n) = n\nu(n)$. Now invoke Fubini's theorem to change the order of integration. This justifies applying bilinearity of the inner product countably many times

$$\begin{aligned}\sum_{s=0}^{\infty} \langle (\mathbf{M}_\gamma)^s u_t, u_t \rangle &= \left\langle \sum_{s=0}^{\infty} (\mathbf{M}_\gamma)^s u_t, u_t \right\rangle \\ &= \langle \mathcal{R}_{\boldsymbol{\gamma}}(1) u_t, u_t \rangle\end{aligned}$$

Because the series is convergent, the operator $\mathcal{R}_{\boldsymbol{\gamma}}(1)$ can be written concisely in $\mathbf{M}_\gamma$: $\mathcal{R}_{\boldsymbol{\gamma}}(1) = (I - \mathbf{M}_\gamma)^{-1}$.

We require the following

Lemma A.0.22 *The operator $\boldsymbol{M}_\gamma$ admits a positive real weak point spectral decomposition*

$\boldsymbol{M}_\gamma = \mathsf{U}\Lambda_\gamma\mathsf{U}'$ *such that every eigenvalue is bounded on unit interval* $1 > m \geq (\Lambda_\gamma)_{j,j} \geq 0$. *Furthermore, at least one eigenvalue is strictly positive.*

Proof The operator $\mathcal{M}'_{\boldsymbol{\gamma}} = \mathcal{M}' - \mu_0\iota'$ has a zero eigenvalue. Moreover, any nonzero eigenvalue ϕ_j has real part strictly bounded inside of the unit interval $[-1+\epsilon_-, 1-\epsilon_+]$, for $1 > \epsilon > 0$, by the Perron-Frobenius theorem. Hence by orthogonality any eigenvalue of $\mathcal{M}_{\boldsymbol{\gamma}}$ is bounded inside the same interval. To see this, note that λ_j, ϕ_j such that $\mathcal{M}\phi_j = \lambda_j$ and $j \neq 0$,

$$\mathcal{M}\phi_j = \iota\mu'_0\phi_j + \mathcal{M}_{\boldsymbol{\gamma}}\phi_j = \mathcal{M}_{\boldsymbol{\gamma}}\phi_j = \lambda_j\phi_j$$

because the dual bases of different eigenspaces are orthogonal by construction (for every $j \neq i$, $\phi'_i\phi_j = 0$; in this case, the long run expected value of ϕ_j is zero, written $\mu'_0\phi_j = 0$). Because $\text{rank}(\mathcal{M}) > 1$, there is at least one nonzero eigenvalue of $\mathcal{M}'_{\boldsymbol{\gamma}}$, λ^*_j, with eigenvectors ψ_j,

$$\mathcal{M}'_{\boldsymbol{\gamma}}\psi_j = \lambda^*_j\psi_j \quad \lambda^*_j \in \mathbb{C} \tag{8.j}$$

Clearly, any ψ_j satisfying 8.j gives

$$\begin{aligned}[\mathcal{M}'_{\boldsymbol{\gamma}}\psi_j]'\mathcal{M}'_{\boldsymbol{\gamma}}\psi_j &= \psi'_j\mathcal{M}_{\boldsymbol{\gamma}}\mathcal{M}'_{\boldsymbol{\gamma}}\psi_j \\ &= \lambda^{*2}_j\psi'_j\psi_j > 0\end{aligned}$$

We conclude that $0 < \lambda^{*2}_j < 1$. Recall $\mathbf{M}_\gamma := \mathcal{M}_{\boldsymbol{\gamma}}\mathcal{M}'_{\boldsymbol{\gamma}}$ and write the above condition

$$\langle\psi'_j, \psi'_j\mathbf{M}'\rangle = \lambda^{*2}_j\langle\psi'_j, \psi'_j\rangle$$

That it is an eigenvalue of $\mathbf{M}_\gamma$ follows from

$$\lambda^{*2}_j = \frac{\parallel \mathcal{M}'_{\boldsymbol{\gamma}}\psi_j \parallel}{\parallel \psi_j \parallel} \geq \frac{\parallel \mathcal{M}'_{\boldsymbol{\gamma}}\psi \parallel}{\parallel \psi \parallel}$$

for any $\psi \neq \psi_j$ since ψ_j is an eigenvalue of $\mathcal{M}'_{\boldsymbol{\gamma}}$.

□

Now, using $\mathcal{R}_\gamma(1)(I - \mathsf{U}\Lambda_\gamma\mathsf{U}') = I$, and the unitary identity $\mathsf{U}' = \mathsf{U}^{-1}$, we have

$$\begin{aligned}(I - \mathbf{M}_\gamma)^{-1} &= (I - \mathsf{U}\Lambda_\gamma\mathsf{U}')^{-1} \\ &= (\mathsf{U}\mathsf{U}' - \mathsf{U}\Lambda_\gamma\mathsf{U}')^{-1} \\ &= \left(\mathsf{U}[\mathsf{U}' - \Lambda_\gamma\mathsf{U}']\right)^{-1} \\ &= \left(\mathsf{U}[I - \Lambda_\gamma]\mathsf{U}'\right)^{-1} \\ &= \mathsf{U}'\left[I - \Lambda_\gamma\right]^{-1}\mathsf{U}\end{aligned}$$

where we have used $\mathsf{U}' = \mathsf{U}^{-1}$ in the second and fifth equality and that $I - \Lambda_\gamma$ is diagonal.

Define

$$(D_{1-\lambda})_{i,j} := \begin{cases} \frac{1}{1-\lambda_j} & i = j \\ 0 & i \neq j \end{cases}$$

We have shown

$$(I - \mathbf{M}_\gamma)^{-1} = \mathsf{U}' D_{1-\lambda} \mathsf{U}$$

Returning to the resolvent decomposition,

$$\begin{aligned} \mathbb{V}(\phi_t) &= \langle \mathsf{U} D_{1-\lambda} \mathsf{U}' u, u \rangle \\ &= \langle D_{1-\lambda}^{1/2} \mathsf{U}' u, D_{1-\lambda}^{1/2} \mathsf{U}' u \rangle \\ &= \langle D_{1-\lambda}^{1/2} u, D_{1-\lambda}^{1/2} u \rangle \\ &= D_{1-\lambda}^{1/2} \langle u, u \rangle D_{1-\lambda}^{1/2} \\ &= D_{1-\lambda} \Sigma \end{aligned}$$

Write $R \cdot \mu_0 = R_0$ and consider the SVD of the sample fluctuations of realized returns around their mean: $R_{t-T,t} - \overline{R}_T = \mathsf{V} D^{1/2} \mathsf{W}'$. The covariance matrix is simply

$$\begin{aligned} \mathbb{V}(R_t) &:= \langle R_t, R_t \rangle - R_0 R_0 \\ &= \mathsf{V} D^{1/2} \mathsf{W}' \mathsf{W} D^{1/2} \mathsf{V}' \\ &= \mathsf{V} D \mathsf{V}' \end{aligned}$$

Of course, $D = \Lambda_{PCA}$. Hence,

$$\begin{aligned} \mathsf{V} \Lambda_{PCA} \mathsf{V}' &= D_{1-\lambda} C C' \\ \langle \mathsf{V} \Lambda_{PCA}^{1/2}, \mathsf{V} \Lambda_{PCA}^{1/2} \rangle &= \langle \mathsf{U} D_{1-\lambda}^{1/2} u, \mathsf{U} D_{1-\lambda}^{1/2} u \rangle \end{aligned}$$

□

Remark: From the proof of proposition (7.2.1), we can see the unitarily equivalent bundles specified by $\div$ can be summarized by the equalities

$$\begin{aligned} \mathsf{V} \Lambda_{PCA} \mathsf{V}' &= D_{1-\lambda} \mathsf{U} \Sigma \mathsf{U}' \\ &= D_{1-\lambda}^{1/2} \Sigma D_{1-\lambda}^{1/2} \\ &= \mathsf{U}' D_{1-\lambda} \mathsf{U} \Sigma \end{aligned}$$

Martingale Representation

Markov dynamics define a serially uncorrelated mean-zero process $u_{t+1} = r(x_{t+1}) - (\mathcal{M}r)(x_t)$ with boundary $u_0 = 0$. We construct a martingale u_t^* recursively $u_t^* = u_t + u_{t-1}^*$ with boundary $u_0^* = r(x_0)$. Using the definition of u_{t+1} and the operator $(\mathcal{L}r)(x_t) = (\mathcal{M}r)(x_t) - r(x_t)$, the

martingale takes the form

$$u^*_{t+1} = r(x_{t+1}) - \sum_{s=0}^{t} (\mathcal{L}r)(x_{t-s}) \tag{3.2a}$$

The *generator* $\mathcal{L} = \mathcal{M} - I$ takes martingales of $\mathcal{M}$ to its kernel. Because $\mathcal{M}$ has eigenvalues in the closed unit circle, $-\mathcal{L}$ has eigenvalues in $[0, 1)$. For values near zero and $\mathcal{M}\psi_j = \lambda_j(x)\psi_j$,

$$\mathcal{L}\psi_j = -e^{-i\lambda_j(x)}\psi_j$$

is a good first-order approximation. Hence, the spectral gap is equivalently the smallest nonzero eigenvalue of the (negative) generator.

Write $\hat{\zeta}_t^{-1} := \zeta_t^{-1}\boldsymbol{\gamma}\gamma x_{t-1}$. From the construction of the martingale u^*_{t+1} in 3.2a, we see that the cumulative process associated with the changes in expected returns is, unsurprisingly, the expected return itself. Using 3.2a with $\Delta\hat{\mathbb{E}}_{t-1,1} \equiv 0$, we can express the expected return process

$$\mathbb{E}_t[R_{t+1}] = \hat{\zeta}_t^{-1}(x_t) - \sum_{n=0}^{t-1} (\mathcal{L}\hat{\zeta}_{t-1-n}^{-1})(x_{t-1-n}) \tag{3.2b}$$

in terms of first differences of the inverse spectral gap.

Covariance Matrix with Non-Centered SVD

The noncentered SVD puts $\phi = A\Lambda_1 B'$. Then,

$$E[\phi - E\phi][\phi - E\phi]' = E\left[A\Lambda_1 B'(I - \tfrac{1}{T}11')\left(A\Lambda_1 B'(I - \tfrac{1}{T}11')\right)'\right] = E[A\Lambda_1 B' I_1 I_1' B\Lambda A']$$

The matrix $I_1 := I - \frac{1}{T}11'$ is symmetric, hence $I_1 I_1' = (I_1)^2$. Element-wise,

$$\begin{aligned}(I_1)^2_{i,i} &= (1 - T^{-1})^2 + T^{-2}(T-1) = 1 - T^{-1} = (I_1)_{i,i}, \text{and,}\\ (I_1)^2_{i,j} &= -2(1 - T^{-1})T^{-1} + T^{-2}(T-2) = -T^{-1} = (I_1)_{i,j}\end{aligned}$$

so I_1 is a projection. Hence $E_T[A\Lambda_1 B' I_1 I_1' B\Lambda A'] = T^{-1} A\Lambda_1 B' I^{[T]} B\Lambda A'$ where $I^{[T]} = TI_{T\times T} - 1_T 1_T' = TI_1$.

Under mild restrictions,

$$T^{-1} A\Lambda_1 B' I^{[T]} B\Lambda A' \longrightarrow_{T\to\infty} \mathbb{V}(\phi)$$

where $\mathbb{V}(\phi)$ is the population covariance of ϕ.

A.0.6 Changes of Measure using Deviations from μ_0

The process has long run mean μ_0. We calculate the empirical distribution $\widehat{h}_t = \frac{1}{T}h_t$ where $h_t = \sum_{s=1}^{T} \delta_x(s)$ which is different than μ_0 in almost every finite sample. Precisely, for $a_\epsilon := \mu_0 + \epsilon$,

$$\begin{aligned}\mathbb{P}(h_t \geq T a_\epsilon) &\leq e^{-kTa_\epsilon}\mathbb{E}[e^{k\,h_t}] \\ &= e^{-TI(a_\epsilon)-o(T^{-1})}\end{aligned}$$

where $I(a_\epsilon) = \sup_k\{ka_\epsilon - \log(\mathbb{E}[e^{k\,h_t}])\}$.

Remark: In the standard normal case, optimality puts $k = \epsilon$. Hence $I(x) = \frac{1}{2}x^2$ and $I(0) = I'(0) = 0$. $I(0) = 0$ expresses the law of large numbers, while $I'(0) = 0$ captures the inflection point of $I(x)$, which locally measures the rate at which the large-time outcomes can deviate from their limiting behavior. $I(x)$ is dubbed the *rate function*; see Veradhan (1979, 2008).

Proofs of Analytic Lemmas

Proof of Lemma 7.2 For each $f \in F(X)$, the Riesz representation theorem identifies the linear functional

$$\eta(f) = \int f\eta_f = \langle f, \eta_f \rangle$$

as the inner product of f against a unique function $\eta_f \in F(X)$. We can take $\eta(f) = (\mathcal{M}f)(1)$ for consistency, but all linear functionals take the form needed.

First consider the simple functions $f = \sum_i 1_{E_i}$, for disjoint E_i with $\bigvee_i E_i = (2^X)^{\mathbb{N}}$. Using the countable additivity of the Stieltjes integral, clearly

$$(\eta(f))(E_i) = \int 1_{E_i}\eta_f = \mathbb{P}_{\eta(f)}(E_i)$$

is a countably additive set function. Because the path space is countable, the integral is finite on every subset of $(2^X)^{\mathbb{N}}$ and trivially regular (Tao 2010, p. 152, 1.10.12). We conclude by Theorem 1.10.11. in Tao (2010) that for each i, $\mathbb{P}_{\eta(f)}(E_i)$ is a Radon measure. Moreover, $\eta(f)(Z) = \sum_i \mathbb{P}(E_i) = 1$ for any disjoint partition of $(2^X)^{\mathbb{N}}$ so $\mathbb{P}_{\eta(f)}$ is a probability measure on $(X^{\mathbb{N}}, (2^X)^{\mathbb{N}})$. We conclude each simple function $f \in F(1_X) \subset F(X)$ corresponds to a unique $\eta_f \in F(1_X)^*$ which itself corresponds to a probability measure $\mathbb{P}_{\eta(f)}$ (uniquely up to functions that agree a.e.).

Recall the definition of the adjoint map $\mathcal{M}'$ given $\mathcal{M}$,

$$\langle \mathcal{M}f, g \rangle = \langle f, \mathcal{M}'g \rangle$$

with $g = \eta_{\mathcal{M}f} \in F(X')^*$ and hence $\mathcal{M}'g = \mathcal{M}'\eta_{\mathcal{M}f} \in F(X)^*$. The adjoint exists by the Riesz representation theorem (Tao (2010), p. 54, 1.4.15) and (if necessary) extends to all of X^* by the Hahn-Banach theorem. In particular, the adjoint map can be constructed directly for each

$\eta_f \in F(X')^*$ via the composition

$$\mathcal{M}'\eta_f = \eta_f \circ \mathcal{M}$$

(Rudin 1991, p. 98 Theorem 4.10). We conclude that for simple functions $f \in F(1_X)$, the adjoint operator $\mathcal{M}'$ maps the dual of the range of $\mathcal{M}$ to the dual of the range of $\mathcal{M}^{-1}\mathcal{M}$, (= I on Hilbert space), and we can identify each element of $F(1'_X)^*$ with a probability measure over $\mathscr{R}(\mathcal{M})$.

The collection of simple functions $F(1_X)$ is dense in $F(X)$. By the Stone-Weierstrass theorem, $F(X)$ can be obtained from $F(1_X)$ by including the limit points of $F(1_X)$. Now, applying the linear functionals η_f to the limit points of $F(1_X)$, it is clear that under the weak* topology the collection of probability measures corresponding to simple functions is dense in the space of probability measures over $\mathscr{R}(\mathcal{M})$. The corresponding limit points are obtained from the weak* limits of functionals of simple functions

$$\eta(f_n) \longrightarrow \eta(f_\infty)$$

The result follows by application of the Stone-Weierstrass theorem to the dual $F(X')^*$ given $F(1'_X)^*$. □

Following Tao (2010, p.14), the Reisz functional representation $\eta(f) = \int f \mathrm{d}(\eta_f)$ defines a measure m such that $\int \mathrm{d}\eta_f = \int f \mathrm{d}m$ and for any g, $\int g\eta_f = \int gf \mathrm{d}m$.

Proof of lemma 7.7b Recall that $g(\{0\}) = 1$ implies that the subspace

$$E_0 := \{\hat{\mu} : (\mathcal{M}' - \lambda_0 I)\hat{\mu} = \mathbf{0}\}$$

has dimension one. Any $\hat{\mu}$ such that $\mathcal{M}'\hat{\mu} = \hat{\mu}$ has the form $\hat{\mu} = c\mu_0$ for non-zero $c \in \mathbb{R}/\{0\}$. Hence $\mathrm{rank}(\mu_0\iota') = 1$, and in this case, $c \equiv 1$ for every row of $\mathcal{M}$. Moreover, $\dim \mathrm{Span}(I_S) = S$. By the rank-nullity theorem, the kernel of $\mu_0\iota'$ has dimension $S - 1$. Because $\mu_0\iota' \perp I - \mu_0\iota'$, the range of $I - \mu_0\iota'$ coincides with the nullspace of $\mu_0\iota'$. Hence $\mathrm{rank}(I - \mu_0\iota') + \mathrm{rank}(\mu_0\iota') = S$. In finite dimensions, all bases are isometrically isomorphic, so if the rank of two operators are equal, their span is equivalent (up to unitary maps). We conclude

$$\{v : v = \mathcal{M}'u\,,\, u \in \mathbb{R}^S\} = \left\{v = v_0 + v_1 : v_0 = \mu_0\iota' u\,,\, v_1 = (I - \mu_0\iota')u\,,\, u \in \mathbb{R}^S\right\}$$

Together, $\mu_0\iota', (I - \mu_0\iota')$ span the image of X under $\mathcal{M}'$.

□

Remark: The characteristic polynomial of $\mathcal{M}' - I$ can be expressed

$$\det(\mathcal{M}' - \lambda I) = (\lambda - \lambda_0)^{\chi(0)} \prod_{j=1}^{N_0 - 1} (\lambda - \lambda_j)^{\chi(j)}$$

with the algebraic multiplicity $\chi(0) = 1$. The characteristic polynomial can also be written

$$\det(\mathcal{M}' - \lambda I) = \left(\lambda - \iota'\mathcal{M}'\mu_0\right)^{\chi(0)} \prod_{j=1}^{N_0 - 1} \left(\lambda - v_j'\mathcal{M}'\nu_j\right)^{\chi(j)}$$

Hence for $\mathrm{rank}(\mathcal{M}) > 1$, the Perron-Frobenius theorem implies that for some $\lambda_1 < 1$,

$$v_1'\mathcal{M}'\nu_1 = \lambda_1$$

to ensure the additional singularity outside a neighborhood of $\lambda_0 = 1$.

Appendix B

Model Development and Proofs

B.1 Benchmark Results

Here we state and discuss implications from the complete and incomplete markets benchmark models. We present the results necessary for the comparisons referenced in the body of the paper, including results on welfare, asset pricing and securities positions, as well as the ex-post wealth distributions. Some additional results not central to the comparisons above are listed along with the proofs in section A.1.

B.1.1 Complete Markets

To complete the markets, we include a perfectly enforced Arrow-Debreu contingent claim $a(s_{m,k}, j)$ for every distinct state of the economy $(s_{m,k}, j)$, $m = R, G$, $k = A, B$, $j = 1, 2$. Details of the trading technology are provided in section 6.0.1.

Proposition B.1.1 (Complete Markets Benchmark) *In the complete markets economy with ex-ante identical investors and security menu S that spans consumption paths*

1. *The log-utility representative agent pricing kernel $M[(s_{m,k})]$ is constructed state-by-state*

$$\nu_0 \, M[(s_{m,k})] = \overline{Y}_{m,k}^{-1} \, \pi_{m,k}$$

for each $s_{m,k} \in S$ and time-zero marginal value of wealth ν_0.

2. *Efficient allocations can be implemented with assets from the benchmark economy 1*
3. *The ex-post distribution of wealth is degenerate.*

The complete markets implications are well known. In section 6.1 of Appendix A, we give a constructive proof that provides the equilibrium contracts $a^*(s_{mk,j})$ in terms of the assets $(a_0, a_j(n_{1,j}, m))$ from section 1.1. We state the key constructions here:

Supporting Positions A standard argument, reproduced in 6.1.2, shows the equilibrium complete markets wealth shares are identical $\theta_{T,j}(s_{m,k}) = \theta_T(s_{m,k}) = \overline{Y}_{m,k}$ for all individuals $j \in \mathcal{I}$. Securities positions a_j supporting the risk sharing rule are

$$a(s_{m,k}, n_j)_{\text{Complete}} =: a_j^0 = -n_{1,j}[Y_{m,k} - V_{1,m}]^{-1} \tag{S.0.1}$$

for every $s_{m,k} \in S$. Supporting positions are derived in section 6.1.3.

Asset Prices From the equilibrium wealth shares $\theta_{T,j}(s_{m,k})$, any investor's marginal value of wealth recovers the representative Lucas pricing kernel state-by-state

$$\frac{d}{d\theta}\log[\theta_{T,j}(s_{m,k})]\pi_{m,k} = [\theta_T(s_{m,k})]^{-1}\pi_{m,k} = \nu_0\, M[\overline{Y}_{m,k}]_{\text{Lucas}}$$

for each $s_{m,k}$. The full system of state prices for $s_{m,k} \in S$ are obtained in this way. In particular, $q(s_{m,k}, n_j) = q(s_{m,k})\pi_j = \nu_0^{-1}[\theta_{T,j}(s_{m,k})]^{-1}\pi_{m,k}\frac{1}{2} = \nu_0^{-1}\frac{1}{2}[\overline{Y}_{m,k}]^{-1}\pi_{m,k} = \frac{1}{2}M[\overline{Y}_{m,k}]_{\text{Lucas}}$.

B.1.2 Incomplete Markets

The set-up is identical to the baseline case presented in section 1.[1] The objective is restated with the proofs in section 6.2 of Appendix A.

[1] Program 1.A is optimized and markets 1.C clear.

Proposition B.1.2 (Incomplete Markets Benchmark) *For the incomplete markets model 5.2, an equilibrium exhibits the following properties*

1. *There is a no trade equilibrium at time $t = 0$.*
2. *Trade at $t = 1$ induces a distribution of wealth $\theta_{T,j}(s_{m,k})$ with shares for each investor j proportional to her realized $n_{1,j} \in \{-\Delta, \Delta\}$.*
3. *The incomplete markets pricing kernel $M[\theta_{t,j}(s_{m,k})]_{NC}$ can be written in terms of the complete markets kernel and the wealth distribution state-by-state*

$$M[\theta_{T,j}(s_{m,k})]_{NC} = \left[M[\overline{Y}_{m,k}]_{Lucas}\right] \; e^{\sigma_\Delta(s_m)(1+\frac{1}{2}\sigma_\Delta(s_m))-o(\frac{1}{\Delta})}$$

for $s_{m,k} \in S$, and where $\sigma_\Delta(s_m) = \left(\frac{\Delta}{W_{1,m}(s_m)}\right)^2$ and

$$M[\overline{Y}_{m,k}]_{Lucas} = \left[\frac{\partial J_1}{\nu_0}\,\pi_m\right] \; \frac{\overline{Y}_{m,k}^{-1}}{\partial J_1}\,\pi_k$$

   (a) *Prices are strictly higher than complete markets prices. The difference is proportional to the welfare loss.*

Remark The cross-sectional transfers $\{n_{t,j}\}_{j\in\mathcal{I}}$ stimulate rebalancing activities. At time-zero, expected marginal values are distorted by uncertainty about $n_{t,j}$. At time 1, the distribution of wealth is bimodal, with relatively poor and rich populations corresponding to realizations $-\Delta$ and Δ, respectively.

Remark Complete markets S.0.1 represent $(s_{m,k}, n_{1,j})$ - contingent payments. For example, policy a_j^0 for type j: $n_{1,j} = \Delta > 0$ requires payment to type $-j$: $n_{1,-j} = -\Delta < 0$ in the *low* productivity state, $Y_{m,k} - V_{1,m} < 0$. In contrast, for *any* aggregate state, incomplete markets policies a_j respond to $n_{1,j} = \Delta > 0$ with acquisitions, while policies a_{-j} respond to $n_{1,-j} = -\Delta < 0$ with liquidations.

Proof We first show that at time-zero, no-trade is an equilibrium. We then open markets in response to the realizations $n_{1,j}$ to derive wealth shares, securities positions and prices. Details omitted from the main text are found in section 6.2 in Appendix A

No Trade Given that there is no ex-ante heterogeneity, no trade at time-zero can be seen by assuming every investor consumes her endowment, then using the corresponding (IMRS) as a price system. The proof is given in section 6.2.1.

We now state the wealth shares and describe the supporting securities positions. Asset prices are developed in the following section, 2.2.1.

Risk Sharing Using $\partial J_{1,j} = \frac{\partial}{\partial a_j} J_{1,j}$ and $\overline{Y}_{m,k} = Y_{m,k} + \omega_0$, wealth shares are written

$$\theta_{2,j} = \frac{\partial J_{1,-j}}{\partial J_{1,j} + \partial J_{1,-j}} \overline{Y}_{m,k} \tag{I.θ}$$

The derivation of equilibrium wealth shares is in section 6.2.2 of Appendix A.

Securities Market clearing gives $a_j = -a_{-j}$. Write the wealth shares $\theta_{2,j} = a_j(Y_{m,k} - V_{1,m}) + n_{1,j} + \theta_0$ with common term $\theta_0 := \overline{Y}_{m,k} + a_0(Y_{m,k} - V_0)$. Put $W_{1,m} := V_{1,m} + \omega_0$. For log utility, $\partial J_{1,j} = [W_{1,m} + n_{1,j}]^{-1}$. Using I.$\theta$, these imply $[a_j - n_{1,j} W_{1,m}^{-1}][Y_{m,k} - V_{1,m}] = 0$.

We express the reallocation policies a_j as a partition of $n_{1,j}$ into two components. One component corresponds to adjustments in the equity position via $a_j V_{1,m}$, while the other component maps to "cash." An implementation of the policy a_j can be written, in units of wealth,

$$\begin{aligned} \textbf{Equity:} \quad & a_j V_{1,m} = n_{1,j} \frac{V_{1,m}}{W_{1,m}} = n_{1,j}\, \alpha_1 \\ \textbf{Cash:} \quad & n_{1,j} - a_j V_{1,m} = n_{1,j}[1 - \alpha_1] \end{aligned} \tag{S.1.2}$$

for $n_{1,1} = \Delta$, $n_{1,2} = -\Delta$ and every $s_{m,k} \in S$, and where α_1 is the fraction of aggregate wealth held in the risky asset in equilibrium.

□

Remark The partition in S.1.2 captures the myopic policy formation characteristic of log-utility populations. The response to $n_{1,j}$ simply splits the gain or loss into risky and risk-free components at the same rate that portfolios hold equity and cash in equilibrium.

B.2 Proofs

Recursion

We show indifference to initial aggregate wealth $W_0 > 0$ and future levels of aggregate wealth $W_t > 0$ $t = 1, T$. Standard arguments are used, based on homothetic preferences and properties of log. The additional state variables are discussed.

Proof Recall gross positions as fractions of net worth, α_j, and share adjustments a_j, are connected via $\alpha_j W_{1,j} = (1 + a_0 + a_j)V_{1,m}$, for $W_{1,j} = (1 + a_0)V_{1,m} + \omega_0 + n_{1,j} - a_0 V_0$. Similarly for $t = 0$, $\alpha_0 W_0 = (1 + a_0)V_0$. Using wealth shares α define the returns to wealth over periods $0 \to 1$ and $1 \to 2$, respectively,

$$R_{0,j} := R(a_0, n_j; V_{1,m}) = \alpha_0 \frac{V_{1,m}}{V_0} + (1 - \alpha_0) + \frac{n_{1,j}}{W_0}$$
$$R_{1,j} := R(a_j; Y_{m,k}) = \alpha_j \frac{Y_{m,k}}{V_{1,m}} + 1 - \alpha_j$$

Now, write the wealth process $(W_{1,j}, W_{2,j})$ in terms of W_0 in the natural way. Set $W_{1,j} = W_0 R_{0,j}$ and then $W_{2,j} = W_{1,j} R_{1,j} = W_0 R_{0,j} R_{1,j}$. For convenience, denote the gross return on initial wealth $\theta_{T,j} := W_{2,j}$.

Log utility decouples today's allocation policies from cumulative effects of future policies. Together with the tower property of conditional expectations, and using $n_{1,j}$ *i.i.d.*, it is clear the objective from program 1.A can be written

$$\max_{\boldsymbol{a},\theta} \mathbb{E}_0[\log(\theta_{T,j})] = \log(W_0) + \max_{a_0} \mathbb{E}_0[\log(R_{0,j})] + \mathbb{E}_j \left[\max_{a_j} \mathbb{E}_1[\log(R_{1,j})] \right] \tag{1.B.1}$$

The level $\log(W_0)$ is irrelevant for allocation decisions and therefore irrelevant for asset pricing, from 1.B.1. Moreover, expected utility is unique only up to order preserving transformations, so we can remove the scale factor $\log(W_0)$. Equivalently, without loss of generality,

set $W_0 = 1$.

From 1.B.1 (and 1.B below) we can also disregard future levels of aggregate wealth W_t, $t \leq T$ because log investors require only single-period gross returns $R_{t,j}$ for allocation decisions. Moreover, there is no intermediate consumption. State prices are constructed from shadow values of one-period gross returns.

□

Write the state vector for every individual

$$X_{j,t} := \begin{bmatrix} q_{t,j} \\ V_t \\ T-t \end{bmatrix} \in \mathbb{R}^2_{++} \times \{2, 1, 0\} \tag{X.1}$$

for strictly positive prices $(q_{t,j}, V_t)$. At $t = 0$, the normalization $W_0 \equiv 1$ implies $q_0 = 1$ for every investor.

Define $J(q_{t,j}, V_t, T-t) := \mathbb{E}_t[\log(\Pi_{s=1}^{T-t} R_s(\boldsymbol{a}^*))]$ along the optimal policy path $\boldsymbol{a}^*$. The additional index in $J(q_t; V_t, T-t)$ monitors the number of periods prior to termination, $T-t$, although we adopt the conventional shorthand $J_0 = J(X_0) = J(1; V_0, 2)$ and $J_{1,j} = J(X_{j,1}) = J(q_j; V_{1,m}, 1)$. Indirect utility separates recursively

$$\begin{aligned} J(X_0) &= \max_{a_0} \mathbb{E}_0 \left[\log(R_{0,j})\right] + \mathbb{E}_j \left[J(X_{j,1})\right] \\ J(X_{j,1}) &= \max_{a_j} \mathbb{E}_1 \left[\log\left(R_{1,j}\right)\right] \end{aligned} \tag{1.B}$$

Heterogeneity is tracked by treating the ratio of individual wealth to initial wealth $q_{t,j} W_0^{-1} = q_{t,j}$ as a state variable for each individual. This is equivalent to treating private income $n_{1,j}$ as the individual state following from the fact that, contemporaneously, $q_{1,j} = W_{1,m,j} = W_{1,m} + n_{1,j}$. The uninsurable shock $n_{1,j}$ is necessary conditioning information. Policies satisfying 1.A or 1.B / 1.B.1 are made contingent on type $j \in \{1, 2\}$ for $t \neq 0$.

Finally, policies are made contingent on aggregate prices. The deterministic probabilities $\pi_{m,k,j} = \pi_m \pi_k \frac{1}{2}$ are common knowledge. We conclude the vector $X_{j,t}$ in X.1 is a sufficient statistic for the state of the economy.

Note that prices are $V_t = V([s_{m,k}]_t, T-t, (q_{t,j}, q_{t,-j}))$ for $[s_{m,k}]_1 = s_m$, $[s_{m,k}]_2 = s_{m,k}$ and $[s_{m,k}]_0 = \text{null}$. The endogenous state can be altered in several ways, through a change of variables, and still produce a valid description of the economy.

B.2.1 Proposition 2.1

Back to Section 2.1

Market Arrangements We complete securities markets by including an Arrow-Debreu contingent claim $a(s_{mk,j})$ for each $s_{mk,j} \in S := \mathcal{Y} \times (n_{1,1}, n_{1,2})$, the set of all pairs $(Y_{m,k}, n_{1,j})$. For example $s_{GB,2} = (Y_{G,B}, n_{1,2})$. Each claim is traded at time $t = 0$. Contracts are fully enforceable. Arrow-Debreu prices are $q(s_{m,k}, n_j)$. We nest the positive supply endowments e_0 into the contingent claims menu S. Market clearing for every $s_{m,k}$ is $\sum_j a(s_{mk,j})\pi_j = \frac{1}{2}a(s_{m,k}, n_{1,1}) + \frac{1}{2}a(s_{m,k}, n_{1,2}) = \overline{Y}_{m,k}$.

The Economy

The present value of all expenditures net of endowments must equal zero. Write the objective

$$
\begin{aligned}
J_0(q_0; V) \;&=\; \max_{\boldsymbol{a},\theta} \mathbb{E}\left[\log(W_{1,j})\right] \qquad (2.0)\\
\text{s.t.} \qquad & \sum_k \sum_m q(s_{m,k}) \sum_{j=1,2} \frac{1}{2} a(s_{m,k}, n_j)\pi_{m,k} \leq \omega_0 + V_0\\
& \theta_{1,j}(s_{m,k}) = W_{1,j}
\end{aligned}
$$

where now a single time zero budget constraint includes the complete set of marketable securities spanning aggregate and individual j-shocks. Shares of equity are fixed by the endowments giving $\mathbb{E}^Q[e_0] = V_0$. $\theta_{1,j}(s_{m,k}) = \theta_{t=1}(s_{m,k}, n_j)$ is the gross return to initial wealth $W_0 \equiv 1$ in state $(s_{m,k}, n_{1,j})$.[2] For details of the securitization of $n_{1,j}$ see the Securitization section below.

We can read off the first order conditions by inspection

$$-\nu_0\, q(s_{m,k}) + [\theta_{1,j}(s_{m,k})]^{-1}\pi_{m,k} = 0 \tag{5.a}$$

for every $s_{mk,j}$, where $\nu_0 = J_0'(q_0, V)$ is the initial marginal value of wealth, and $q(s_{m,k})$ is the price of a claim to one dollar in state $s_{m,k}$. ν_0 is necessarily identical across investors. Write $\lambda_{0,j}$ for the Lagrange multiplier on the initial budget constraint for investor $j \in \mathcal{I}$. The envelope is $\lambda_{0,j} = J_{0,j}'(q_0, V)$. By ex-ante symmetry, $\lambda_{0,j} = \lambda_0$ for $j \in \mathcal{I}$, so $J_{0,j}' = J_0' = \nu_0$.

Wealth Shares

Proof From equation 5.a, for each $s_{m,k} \in S$ and any two $i, j \in \{1, 2\}$, market prices enforce

$$\theta_{T,i}(s_{m,k}) = \theta_{T,j}(s_{m,k})$$

Investors all have identical final wealth shares $\theta_{T,j}$, which must in turn equal the average share and the total level

$$\theta_{T,i}(s_{mk,i}) = \frac{1}{2}\sum_j \theta_{T,j}(s_{mk,j}) = \omega_0 + Y_{m,k}$$

where the second equality follows by market clearing for terminal wealth. In particular, for every infinitesimal investor and for all $s_{mk,j} \in S$, $\theta_{T,j}(s_{mk,j}) = \omega_0 + Y_{m,k} = \overline{Y}_{m,k}$ is the

[2] θ is a control dummy for final wealth, used for convenience. If a reader prefers to think of utility defined over consumption of a non-perishable numeraire good at the terminal date, written say, $c_1(s_{m,k}, n_j)$, then $\theta_{1,j}(s_{m,k}, n_j) = c_1(s_{m,k}, n_j)$.

complete markets (perfect) risk sharing rule.[3] By populations j, for $\pi_j = 1/2$, the rule is $\widehat{\theta_1}(s_{k,h}, n_j) = \frac{1}{2}[\omega_0 + Y_{m,k}]$.

□

Back to Section 2.1

Supporting Positions

Proof We implement the complete markets risk sharing rule using assets from the benchmark economy. We use a completed menu (containing j-contingencies) of the two-step assets described in the binomial model.[4] Individual portfolio realizations for each $s_{mk,j}$, can be written

$$\theta_{1,i}(s_{mk,i}) = \omega_0 + A_j(s_{mk,j}) + n_{1,j}$$

We recover the allocations by unpacking

$$A_j(s_{mk,j}) = (1 + a_0 + a_{j,m})Y_{m,k} - a_0 V_0 - a_{j,m} V_{1,m}$$

Using $\theta_{1,i}(s_{mk,i}) = \theta_{1,j}(s_{mk,j})$ and $a_0 = 0$, simple algebra reveals

$$[a_{j,m} - a_{-j,m}][Y_{m,k} - V_{1,m}] = n_{1,-j} - n_{1,j}$$

Finally, we appeal to scarce resources $\frac{1}{2}\sum_j (1 + a_{j,m}) = 1$, having used $a_0 = 0$. Recalling that $n_{1,j} + n_{1,-j} = 0$, the remaining allocations can be expressed, in terms of equilibrium

[3]It is straightforward that when net private income distributions $\sum_j \pi_j \Delta_j =: \overline{\Delta} \neq 0$, shares by type j are $\pi_j \left[\omega_0 + \overline{\Delta} + Y_{m,k}\right]$.

[4]These assets are more useful for analyzing the different welfare implications across the three economies we consider. Unsurprisingly, allocations using (a_0, a_j) are equivalent for asset pricing and welfare analyses to allocations using the Arrow-Debreu menu $\mathbf{a}_0 := a(s_{m,k}, n_j)_{\{m,k,j\}}$. Formalities are addressed in the Equivalence discussion of this Appendix.

objects,[5]

$$a(s_{m,k}, n_j)_{\text{Complete}} =: a_j^0 = -n_{1,j}[Y_{m,k} - V_{1,m}]^{-1} \tag{S.0.1}$$

for every $s_{m,k} \in S$.

□

Asset Prices Given the symmetric wealth shares $\theta_{2,j}(s_{m,k})$, any investor's marginal value of wealth can be written in terms of aggregates (identically) and used to price assets. See the Asset Pricing discussion in Section 2.1.

Securitization We can write $\mathbb{E}^Q[e_0] = V_0$ for the unit price of market equity. Note that $\mathbb{E}^P[n_{t,j}] = 0$ for $dQ = e^{-\eta(s)}dP$ but $\mathbb{E}^P[e^{-\eta(s)}n_{t,j}]$ is an equilibrium object.[6] There are several ways to allow $n_{t,j}$ to be marketable. We adopt the simplest case for the present-value representation of our economy by securitizing claims to $n_{t,j}$ at time zero. The equilibrium value for a claim to n_j is

$$\mathbb{E}^Q[n_{t,j}] = \sum_{k=U,D} \sum_{m=R,G} q(s_{m,k}) \sum_{j=1,2} \frac{n_j}{2} \pi_{m,k} = 0 \tag{1n}$$

Securitizaion of n_j has no impact on the level of tradeable wealth at time-zero W_0.[7]

---

[5] In terms of model primitives

$$a(s_{m,k}, n_j)_{\text{Complete}} = -n_{1,j} \left[Y_{m,k} - \mathbb{E}_1 \left[Y_{m,k}[Y_{m,k} + \omega_0]^{-1} \right] \right]^{-1}$$

for every $s_{m,k} \in S$ and where $\mathbb{E}_1 \left[Y_{m,k}[Y_{m,k} + \omega_0]^{-1} \right] = V_{1,m}\nu_0 = \mathbb{E}_1^Q[Y_{m,k}]\nu_0$ with $\nu_0 = 1$ for $u(W_2) = \log(W_2)$ and $W_0 \equiv 1$.

[6] When $n_{t,j}$ is orthogonal to the pricing kernel, $\mathbb{E}^Q[n_{t,j}] = 0$. In complete markets with private shocks that are aggregate-neutral this condition is satisfied. The process $\eta(s)$ has $\mathbb{E}^P[e^{-\eta(s)}] = 1$.

[7] Of course, the tradeability of $n_{t,j}$ shows up as an additional lever in the allocation policies $a = a(s_{m,k}, n_j)$.

Budget Constraint We use the resource restriction

$$\sum_j \frac{1}{2} a(s_{m,k}, n_j) = Y(s_{m,k}) \qquad s_{m,k} \in S$$

together with 1n to write the investor's complete markets budget constraint

$$\sum_k \sum_m q(s_{m,k}) \sum_j \frac{1}{2} [a(s_{m,k}, n_j) - 2Y_{m,k}] \pi_{m,k} \leq \omega_0$$

which states the present value of all financed positions net of endowments is zero.

□

Back to Section 2.1

Equivalence Unsurprisingly, the allocations $\mathbf{a}_0$ in the time-zero economy are equivalent to the allocations (a_0, a_j) in the original two-step economy. By implementing the complete markets risk-sharing rule with a feasible allocation of assets consistent with the trading protocol from the two-step economy, we have shown that *an* allocation in the two-step economy $(\hat{a_0}, \hat{a}_j)$ is weakly preferred to $\mathbf{a}_0$.[8] Because $\mathbf{a}_0$ is Pareto efficient in a frictionless economy with resources and time-separable preferences that are identical to those in the two-step economy, it must also be that $\mathbf{a}_0$ is weakly preferred to *any* $(\tilde{a}_0, \tilde{a}_j)$. Out of these we pick $(\hat{a_0}, \hat{a}_j)$ and set $(a_0, a_j) = (\hat{a_0}, \hat{a}_j)$.

□

B.2.2 Incomplete Markets: Proposition 2.2

Back to Section 2.2

[8] That is, you would never do worse by optimizing in the two-step economy directly.

The Economy

Each investor faces the objective

$$
\begin{aligned}
J(q_0; V_0) \;=\; & \max_{\boldsymbol{a},\theta} \mathbb{E}\left[\log(\theta_{2,j})\right] && \text{(IC.1)}\\
\text{s.t.} \quad & a_0 V_0 - V_0 \leq \omega_0 \\
& a_j V_{1,m} - (1+a_0) V_{1,m} \leq (\omega_0 - a_0 V_0) + n_{1,j} \\
& \theta_{2,j} = \overline{a}_j Y_{m,k} + (\omega_0 + n_{1,j} - a_0 V_0 - a_j V_{1,m})
\end{aligned}
$$

where $\overline{a}_j = (1 + a_0 + a_j)$ and we distinguish the final portfolio value $\theta_{2,j} = W_{2,j}$ from the wealth process $W_0, W_{1,j}, W_{2,j}$. Note that while $n_{t,j}$ cannot be securitized, after $n_{1,j}$ is realized all wealth is tradeable.

Definition: Incomplete Markets Equilibrium In equilibrium, every investor optimizes IC.1 and markets clear according to 1.C.

No Trade Equilibrium

Proof Endowments and preferences are identical. Suppose a price system at time-zero for aggregate states $s_{m,k} \in S$ is given by

$$
\begin{aligned}
q(s_{m,k})\nu_0 =& [\frac{1}{2}[\hat{e_0} + n_{1,1}]^{-1} + \frac{1}{2}[\hat{e_0} + n_{1,2}]^{-1}]\pi_{m,k} \\
=& \left[[\hat{e_0} + \Delta]^{-1} + [\hat{e_0} - \Delta]^{-1}\right] \frac{1}{2}\pi_{m,k}
\end{aligned}
$$

where $\hat{e_0} = \hat{e}_0(s_{m,k}, t)$ is the realization claimed by an investor owning e_0 in state $s_{m,k}$ and period t. In the terminal period, $\hat{e_0} = Y_{m,k}$, while in the interim period $t = 1$, $\hat{e}_0$ is the capital value $V_{1,m}$. The gross rate of time discount is $1 + \beta = 1$.

When we propose a no-trade allocation, feasibility is automatic. Every investor holds her

endowment. Moreover, the two-period horizon circumvents the need to verify transversality conditions. We are left to verify optimality.

The investors face the same number of contingencies as in the complete markets case, but they can only access half the number of primitive assets, corresponding to the cardinality of $\{s_{m,k}\}_{m,k}$. Consider an economy with trade and write the wealth shares $\theta_{t,j}(s_{m,k})$ for $t = 1, 2$. For each tradeable contingency $s_{m,k}$, first order conditions are

$$-\nu_0\, q_\theta(s_{m,k}) \; + [\frac{1}{2}[\theta_{t,j}(s_{m,k})]^{-1} + \frac{1}{2}[\theta_{t,j}(s_{m,k})]^{-1}]\pi_{m,k} = 0$$

where $\frac{1}{2} = \pi_j$ is used, and $\nu_0 = J_0'(q_0, V)$ is the initial marginal value of wealth, necessarily identical across investors.

In contrast, we have proposed prices that correspond to the intertemporal tradeoff

$$-\nu_0\, q(s_{m,k}) + \left[[\hat{e_0} + \Delta]^{-1} + [\hat{e_0} - \Delta]^{-1}\right] \frac{1}{2}\pi_{m,k} = 0$$

for every time-zero investor and any state $s_{m,k} \in S$. The only hope for improving this margin is to pick a $\theta_{1,j}$ to reduce the Jensen cost over states j conditioning on $s_{m,k}$. By assumption, there are no securities to trade on the realizations $n_j \in \{-\Delta, \Delta\}$ and hence, $\theta_{1,j}$ is contingent on n_j only through $\theta_{1,j} = \theta_1^* + n_{1,j}$ where θ_1^* is a control variable at time zero. Investors still must average over $n_{1,j}$ realizations for each $s_{m,k}$. Thus, we can take $\theta_{1,j} = \hat{e}_0 + n_{1,j}$ and $q_\theta = q$. The proposed price system is an optimum for every investor, is feasible, and clears markets.

□

Back to Section 2.2

Incomplete Markets Wealth Shares

Every investor has an identical portfolio *coming in* to the first period, prior to realization of shocks $n_{1,j}$. In response to $n_{1,j}$ and the signal m investors enter securities markets to arrange their final portfolios. Outgoing positions take the form

$$W_{1,j} = \underbrace{V_{1,m}a_j + V_{1,m}}_{\text{Equity Claim}} + \underbrace{\omega_0 + n_{1,j} - a_j V_{1,m}}_{\text{Risk Free Holdings}}$$

and differ for each m only through the pairs $(n_{1,j}, a_j)$.

Proof We derive policies a_j by first extracting the risk sharing rules $\theta_{2,j}$. Market prices enforce

$$-V_{1,m}\frac{\partial}{\partial a_j}J_{1,j} + [\theta_{2,j}]^{-1}Y_{m,k}\pi_{m,k} = 0$$

giving the rule

$$\theta_{2,j}\,\frac{\partial}{\partial a_j}J_{1,j} = \theta_{2,-j}\frac{\partial}{\partial a_{-j}}J_{1,-j}$$

Risk sharing is full conditioning on today's uninsured shock, so variation in the wealth shares $\theta_{2,j}$ across $j = 1, 2$ is driven by today's marginal value of wealth. We adopt shorthand $\partial J_{1,j} = \frac{\partial}{\partial a_j}J_{1,j}$ and $\overline{Y}_{m,k} = Y_{m,k} + \omega_0$. Wealth shares

$$\theta_{2,j} = \frac{\partial J_{1,-j}}{\partial J_{1,j} + \partial J_{1,-j}}\overline{Y}_{m,k}$$

follow from final-period goods market clearing by state $s_{m,k}$, written $\theta_{2,j} + \theta_{2,-j} = 2\overline{Y}_{m,k}$.

□

Time-zero Shares and State Prices

First-order conditions for an asset that pays 1 in state s_m are

$$\nu_0\, q(s_m) - \frac{1}{2}\theta_{1,j}^{-1} - \frac{1}{2}\theta_{1,-j}^{-1} = 0$$

Market clearing gives $\sum_j \alpha_j W_{1,m,j}\, \pi_j = V_{1,m}$. Then $W_{1,m} = V_{1,m} + \omega_0$ and $W_{1,m,j} = W_{1,m} + n_{1,j}$. Put $\theta_{1,j} = \partial J_{1,j}^{-1} = W_{1,m,j} = W_{1,m}^* + n_{1,j}$ where the * indicates the component can be controlled from time-zero. Plugging $\theta_{1,j} = \partial J_{1,j}^{-1}$ into the time-one shares and using $\frac{1}{2}\sum_j \theta_{2,j} = \overline{Y}_{1,m}$ gives $V_{1,m}$. Plugging $V_{1,m}$ into time-zero FOCs using $\theta_{1,j} = W_{1,m} + n_{1,j}$ gives

$$\nu_0\, q(s_m) = W_{1,m}\left([W_{1,m} + \Delta][W_{1,m} - \Delta]\right)^{-1}\, \pi_m$$

in agreement with NC.

□

Nonseparable Preferences To preserve the comparison in the previous sections, we define perishable consumption in $t = 1$ to be a small dividend paid by the productive asset that is a constant proportion of the expected payout conditioning on that path

$$c(m;\epsilon) = \epsilon\, \mathbb{E}[Y(s_{m,k})|m]$$

Then, individual consumption policies are written $c = c(j, m)$. Define

$$u(c, \theta_{T,j}) := \left[c^{1-1/\psi} + \theta_{T,j}^{1-1/\psi}\right]^{\psi/(\psi-1)}$$

and

$$U = \frac{1}{1-\gamma} u(q_1, \theta_{T,j})^{1-\gamma}$$

In addition, resources are now constrained in $t = 1$ by

$$\frac{1}{2}\sum_j c(j, m) = c(m;\epsilon)$$

B.3 Empirical Implications and Evidence

Rather than understanding intermediaries as economic actors themselves our theory suggests the bank's balance sheet uniquely captures demand for liquidity in the cross-section of investors. The endogenous structure of the banking arrangement suggests a particular function of the wealth distribution is captured by a particular function of bank balance sheet. This generates testable implications for a "representative" agent asset pricing theory. The theory is doubly productive because it provides tests that exploit data on financial institutions rather than individual-level data on income, net worth, human capital, real estate etc.

The empirical implication of the representative agent prediction is that shocks to the marginal rate of liquidity production, measured by the rate of risky assets to uninsured liquid liabilities, will have cross-sectional pricing power in markets where assets are accessible broadly to both institutions and individuals. Assets relevant to our predictions include exchange traded stocks and indices. Arguably tests are relevant in fixed income and options markets, which exhibit lower participation rates because of decisions not to enter rather than prohibitions.

The theory also predicts that the component of the pricing kernel that bank balance sheet statistics capture uniquely arises from inter-temporal hedging motives of the reference investor.[9] In large incomplete markets economies with institutional liquidity production, myopic investors induce inter-temporal hedging motives in the SDF through changes in an aggregate measure of their propensity to over-save reflected in bank financing flows.[10,11] Moreover, the wealth distribution channel can operate holding aggregate cash flows fixed, which highlights the propensity for intermediary balance sheets to generate pure discount

[9]The term reference is used in place of representative when the marginal investor cannot be recreated from linear combinations of individual investors in the model.

[10]Similarly, Chien, Cole and Lustig (2012) and Chien and Lustig (2010) find i.i.d. dynamics can still produce persistent risk prices in large incomplete markets economies.

[11]This point is elaborated in a dynamic version of this model, available upon request.

rate effects in the time series of asset returns. An interesting implication is that a decomposition of the log pricing kernel in this model using the present-value identity (Campbell and Shiller, 1989) should reflect that the liquidity factor yield does not predict cash flows.

B.3.1 Data

We use data from the Flow of Funds reports by the Federal Reserve for balance sheet information. The data are quarterly, and aggregated sector-wide. We collect data from sectors that finance a significant portion of their assets with demandable liabilities: commercial banks and broker-dealers. We also collect data for off balance sheet asset-backed commercial paper (ABCP) activities reported to the Federal Reserve, a significant fraction of which have bank holding companies or subsidiaries of bank holding companies as their conduits.

We connect repo financing and ABCP-type liabilities to the demandability of deposits. First, repo are provided by many institutional depositors (Gorton) and almost all of them are expected to rollover their financing. Commercial paper also tends to have a large number of buyers, although CP is less often used as a permanent financing policy. The expected perpetuity property of repo is the same as deposits. Empirically, a key difference is that deposits are countercyclical, while repo are pro-cyclical.

Hence, the liabilities used for construction of our series are repurchase agreements, large time deposits, uninsured savings and checkable deposits and ABCP. The exclusion of ABCP appears to have little effect. The inclusion of insured deposits has significant effects on the time-series properties and the cross-sectional exposure patterns of the series. Similarly, repurchase agreements and uninsurable large time deposits are necessary for the series to produce a viable distribution of exposure in the cross-section of equities. The test-asset cross-sections data are from Ken French. Risky assets are measured by corporate equities, mutual

fund shares, and private residential and commercial mortgage-backed securities (MBS). The productivity series is defined as changes in the ratio of risky assets to liquid liabilities. We report the β distributions and the liquidity production time series in the charts below.

Table B.1: Cross Sectional Exposure to Changes in Liability-Side Productivity

	Portfolio	term	estimate	statistic	std.error
1	S1B1	Δ Liquidity	-0.670	-6.678	0.100
5	S1B5	Δ Liquidity	-0.543	-6.731	0.081
6	S2B1	Δ Liquidity	-0.649	-7.577	0.086
10	S2B5	Δ Liquidity	-0.491	-6.808	0.072
11	S3B1	Δ Liquidity	-0.609	-7.881	0.077
15	S3B5	Δ Liquidity	-0.471	-7.206	0.065
16	S4B1	Δ Liquidity	-0.588	-8.641	0.068
20	S4B5	Δ Liquidity	-0.510	-8.134	0.063
21	S5B1	Δ Liquidity	-0.460	-8.806	0.052
25	S5B5	Δ Liquidity	-0.421	-8.287	0.051

(a) The bank balance sheet productivity measures the ratio of high risk assets to liquid liabilities. Comparison of large-small spread and high-low spread (high-low book to market (BTM) ratios). Quarterly balance sheet data for commercial banks and broker-dealers from the Flow of Funds, Board of Governors of the Federal Reserve. We use private depository institutions, issuers of asset-backed securities, and securities brokers and dealers to measure liquidity production. The ratio of high risk assets to liquid liabilities is calculated by classifying liquid liabilities as large time deposits, uninsured checkable and savings deposits, ABCP and repurchase agreements. Risky assets are corporate equities, mutual fund shares, and private residential and commercial mortgage-backed securities (MBS). Monthly Fama -French 3-factor and Carhart model returns data are from 1967 $Q1$ to 2016 $Q4$.

B.4 Organizational Implications: Internal Diversification

In this appendix we present some of the ancillary implications of the theory in more detail.

Figure B.1: Exposures to Changes in Liability-Side Productivity

(a) The distribution of exposure spreads value and size. Both size and value spreads are large. The bank factor should not be identical to value, because the motive to save with large banks appears even for logarithmic investors, and the value premium captures intertemporal hedging demands (i.e., the HML factor is in zero-net supply). Quarterly balance sheet data for commercial banks and broker-dealers from the Flow of Funds, Board of Governors of the Federal Reserve. We use private depository institutions, issuers of asset-backed Securities, and securities brokers and dealers to measure liquidity production.The ratio of high risk assets to liquid liabilities is calculated by classifying liquid liabilities as large time deposits, uninsured checkable and savings deposits, ABCP and repurchase agreements. Inclusion of insured deposits significantly alters the time series. Risky assets are corporate equities, mutual fund shares, and private residential and commercial mortgage-backed securities (MBS). Monthly Fama -French 3-factor and Carhart model returns data are from 1967 $Q1$ to 2016 $Q4$.

Liquidity Production and Asset Diversification

Banks and other intermediaries hold diversified assets on their balance sheets, but typical non-financial public firms do not. Conventional wisdom holds that investors are weakly better off when individual firms concentrate risk in their area of expertise. Does the expertise of financial firms require they hold diversified assets, or are these allocations inefficient?

Financial operations carried out by market makers, broker-dealers, prime brokerages, at IB trading desks, etc, require holding a variety of assets on behalf of clients, or available for trade, or in some cases for risk-management. Expertise in these businesses entails more internal diversification than, for example, a bio-tech start-up. However, notably commercial banks are omitted from this list, yet tend to diversify assets. Moreover, IB balance sheets are often concentrated via emphasis on a small number of issuances.

A separate function of financial institutions provides an alternative explanation for asset diversification. Broker-dealers, dealer banks, commercial banks and bank holding companies effectively use risky assets as inputs for the production of liquidity on their liability-side. This function is carried out optimally when the balance sheet is both diversified and risky.

Corollary B.4.1 (Liquidity Production and Asset Diversification) *Efficiency of liq-*

uidity production increases with asset diversification. In particular, equilibrium liquidity producers hold the market.

Value is created when a liquidity producer can consistently peel off *average returns* from risky investments - say the market returns - and direct them to the subset of investors with the *highest marginal valuation*. Over time this requires calibrating the distribution of asset returns through portfolio choice. It is costly, on average, to concentrate asset risk: competitive markets for liquidity production will drive out otherwise equivalent institutions with higher overall asset volatility. The most efficient liquidity producers will hold the most diversified asset portfolio, all else equal.

Bibliography

[1] Tobias Adrian and Nina Boyarchenko, *Intermediary Leverage Cycles and Financial Stability*, Federal Reserve Bank of N.Y., Staff Report No. 567, August 2013

[2] Tobias Adrian, Erkko Etula and Tyler Muir *Financial Intermediaries and the Cross-Section of Asset Returns*, The Journal of Finance, Vol. LXIX, No. 6, Dec 2014

[3] Douglas W. Diamond and Philip H. Dybvig, *Bank Runs, Deposit Insurance, and Liquidity*, Journal of Political Economy, Vol. 91 No.3 1983

[4] He, Kelley and Manilla, *Intermediary Asset Pricing: New Evidence from Many Asset Classes*, Journal of Financial Economics, vol. 126, issue 1, 2017

[5] Zhiguo He, In Gu Khang and Arvind Krishnamurthy, *Balance Sheet Adjustments during the 2008 Crisis*, IMF Economic Review Vol. 58, No. 1, 2010

[6] Charles J. Jacklin, *Demand Equity and Deposit Insurance*, Stanford GSB Research Paper No.1062, 1987

[7] Joseph G. Haubrich and Robert G. King, *Banking and insurance*, Journal of Monetary Economics 26 (1990) 361-386

[8] Gary Gorton and Andrew Metrick, *Securitized Banking and the Run on Repo*, NBER and Yale SOM Working Paper, 2009

[9] Gary Gorton and Andrew Metrick, *Who Ran Repo?*, NBER and Yale SOM Working Paper, 2012

[10] Anil K. Kayshap, Raghuram Rajan and Jeremy C. Stein, *Banks as Liquidity Providers: An Explanation for the Coexistence of Lending and Deposit-Taking*, Journal of Finance, Vol. LVII, No. 1, Feb. 2002

[11] Robert E. Lucas, *Liquidity and Interest Rates*, Journal of Economic Theory, No.50, 1990

[12] Fernando Alvarez and Urban J. Jermann, *Using Asset Prices to Measure the Persistence of the Marginal Value of Wealth*, 2005, Econometrica, Vol. 73, No. 6

[13] Theodore W. Anderson, *Second Order Moments of a Stationary Markov Chain with Applications*, Technical Report No. 22 U.S. Army Research Office and Stanford, February 1989

[14] Clifford S. Asness, Andrea Frazzini, Ronen Israel and Tobias J. Moskowitz, *Fact Fiction and Value Investing*, Journal of Portfolio Management, Fall 2015

[15] Clifford S. Asness, Andrea Frazzini, Ronen Israel and Tobias J. Moskowitz, *Fact Fiction and Momentum Investing*, Journal of Portfolio Management, Fall 2014

[16] Federico M. Bandi and Andrea Tamoni, *Business Cycle Consumption Risk and Asset Prices*, 2015, Working Paper

[17] Ravi Bansal and Bruce N Lehman, *Growth Optimal Portfolio Restrictions on Asset Pricing Models*, Macroeconomic Dynamics, 1, pp. 333-354, 1997

[18] Ravi Bansal and Amir Yaron, *Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles*, The Journal of Finance, Vol. 59, No. 4 (Aug.), 2004, pp. 1481-1509

[19] Alex Bloemendal Antti Knowles Horng-Tzer Yau Jun Yin, *On the principal components of sample covariance matrices*, Working Paper, Harvard Mathematics Department, February 14, 2015

[20] Michael Brennan and Alex Taylor, *Expected Returns and Risk in the Stock Market*, UCLA Anderson School Working Paper, 2016

[21] Jaraslov Borovicka, Lars Peter Hansen and Jose A Scheinkman, *Misspecified Recovery*, Journal of Finance, Vol.71, No.6, 2016

[22] Svetlana Bryzgalova, *Spurious Factors in Linear Asset Pricing Models*, LSE Working Paper 2014

[23] Svetlana Bryzgalova and Christian Julliard, *The Consumption Risk of Bonds and Stocks*, Working Paper, September 15, 2015

[24] Craig Burnside, *Identification and Inference in Linear Stochastic Discount Factor Models with Excess Returns*, Working Paper, June 2015

[25] Carhart, M. M., *On Persistence in Mutual Fund Performance.*, 1997, Journal of Finance, no 52 pg 57-82

[26] Nai-Fu Chen, Richard Roll and Stephen A. Ross, *Economic Forces and the Stock Market*, Journal of Business, Vol.59 No. 3, 1986

[27] John Cochrane, *Discount Rates*, Presidential Address, American Finance Association, 2011

[28] John Cochrane and Monika Piazzesi, *Bond Risk Premia*, 2005, American Economic Review, v 95 n 1

[29] John Cochrane, *The Dog That Did Not Bark: A Defense of Return Predictability*, 2008, The Review of Financial Studies, v 21 n 4

[30] Persi Diaconis, *The Markov Chain Monte-Carlo Revolution*, Bulletin of the American Mathematical Society, Vol. 46 No. 2 April 2009, pp. 179-205

[31] Persi Diaconis and Laurent Saloff-Coste, *Comparison Theorems for Reversible Markov Chains* The Annals of Applied Probability, Vol.3 No. 8 1993

[32] J.L. Doob, *Discrete Potential Theory and Boundaries*, Journal of Mathematics and Mechanics, Vol.8 No.3, 1959, pp. 433-458

[33] M.D. Donsker and S.R.S. Varadhan, *Asymptotic Evaluation of Certain Markov Process Expectations in Large Time, I*,II*,III**, Communications in Pure and Applied Mathematics, 1975 - 1976, Vol. 28 - 32

[34] Pierre Collin-Dufresne, Michael Johannes, and Lars A. Lochstoer, *Parameter Learning in General Equilibrium: The Asset Pricing Implications*, American Economic Review, 2016

[35] D.Duffie, J. Geanakoplos, A. Mas-Colell, A. McLennan *Stationary Markov Equilibria*, Econometrica, Vol. 62, No. 4. (Jul.), 1994, pp. 745-781.

[36] Lazlo Erdos, Horng-Tzer Yau, *Dynamical approach to random matrix theory*, Institute of Science and Technology, Austria. Manuscript, 2017

[37] Lazlo Erdos, *Notes on the Matrix Dyson Equation*, Lecture Notes, Institute of Advanced Studies, Princeton University, 2016

[38] Eugene Fama and Ken French, *Common Risk Factors in Stocks and Bonds*, Journal of Financial Economics, 33 (1993) 3-56

[39] Eugene F. Fama and Kenneth R. French, *A Five-Factor Asset Pricing Model* September 2014, Working Paper

[40] Gibbons, Ross and Shanken, *A Test of the Efficiency of a Given Portfolio*, Econometrica, Vol. 57, No. 5 (September), 1989, 1121 to 1152

[41] Joseph Gerakos Juhani T. Linnainmaa, *Decomposing Value*, Chicago Working Paper, September 19, 2012

[42] Valentin Haddad, Serhiy Kozak, Shri Santosh, *Predicting Relative Returns*, Working Paper, 2017

[43] Lars Hansen, *Dynamic Valuation Decomposition with Stochastic Economies* NBER and Chicago Working Paper, 2011; Koopmnan and Fisher-Schulz Lecture series, 2008

[44] Lars Peter Hansen and Thomas J. Sargent, *Fragile beliefs and the price of uncertainty*, Journal of Quantitative Economics 1, 2010

[45] Lars Peter Hansen and Jose A. Scheinkman, *Long Term Risk: An Operator Approach*, Econometrica, Vol. 77, No. 1 (January, 2009), 177-234

[46] J. Michael Harrison and David M. Kreps, *Martingales and Arbitrage in Multiperiod Securities Markets*, Journal of Economic Theory, 20, 381-408, 1979

[47] Campbell R. Harvey, Yan Liu, Heqing Zhu, *. . . and the Cross-Section of Expected Returns*, Review of Financial Studies, October 2015

[48] Simon Huang, *The Momentum Gap and Return Predictability*, Working paper, 2016

[49] Ravi Jagannathan; Zhenyu Wang, *The Conditional CAPM and the Cross-Section of Expected Returns* The Journal of Finance, Vol. 51, No. 1. (Mar.), 1996, pp. 3-53.

[50] Brian Kelly, Seth Pruitt, Yinan Su, *Instrumented Principal Components Analysis*, 2017, Working Paper

[51] Brian Kelly and Seth Pruitt, *Market Expectations and the Cross-section of Present Values*, 2013, Journal of Finance, Vol. 68, No. 5

[52] Bryan Kelly, Seth Pruitt and Yinan Su, *Some Characteristics Are Risk Exposures and the Rest Are Irrelevant*, Working Paper, 2017

[53] Peter Lax, *Functional Analysis*, Pure and Applied Mathematics, 2002

[54] Lettau, M., Ludvigson, S.C., *Consumption, aggregate wealth and expected stock returns*, Journal of Finance 56 (3), 2001, 815-849

[55] Thomas M. Liggett, *Continuous Time Markov Processes*, American Mathematical Society, Graduate Texts in Mathematics, Vol. 133, 2010

[56] Andrew W. Lo, *The Statistics of Sharpe Ratios*, Financial Analysts Journal, July/August 2002

[57] David G. Luenberger, *Optimization by Vector Space Methods*, Wiley Professional Paperback Series, 1969

[58] Paulo Maio and Pedro Santa-Clara, *Dividend Yields, Dividend Growth, and Return Predictability in the Cross Section of Stocks*, Journal of Financial and Quantitative Analysis, Vol. 50, Nos. 1/2, Feb./Apr. 2015

[59] Markowitz, Harry, *Portfolio Selection*, Journal of Finance 7, 1952, 77 to 91.

[60] Chistopher J. Malloy, Tobias J. Moskowitz, and Annette Vissing-Jorgensen, *Long-Run Stockholder Consumption Risk and Asset Returns*, Journal of Finance, December 2009

[61] Merton, Robert C., *Optimal Consumption and Portfolio Rules in a Continuous Time Model*, Journal of Economic Theory 3, 1971, 373 to 413.

[62] Merton, Robert C., *An Intertemporal Capital Asset Pricing Model*, Econometrica 41, 1973, 867 to 887.

[63] Michaud, Robert O., *The Markowitz optimization Enigma: Is Optimized Optimal?*, Financial Analysts Journal, Jan/Feb 1989

[64] Alan Moreira and Tyler Muir, *Volatility-Managed Portfolios*, Journal of Finance, May 2017

[65] Tobias J. Moskowitz and Mark Grinblatt *Do Industries Explain Momentum?*, The Journal of Finance, Vol. 54, No. 4, (Aug.), 1999

[66] John Nash, *Continuity of Solutions of Parabolic and Elliptic Differential Equations*, American Journal of Mathematics, Vol.80, No.4, 1958, pp. 931-954

[67] Newey, W.K., West, K.D., 1987. *A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix*, Econometrica 55, 703-708.

[68] Fulvio Ortu, Andrea Tamoni and Claudio Tebaldi, *Long-Run Risk and the Persistence of Consumption Shocks*, The Review of Financial Studies, Vol. 26, No. 11 (November), 2013, pp. 2876-2915

[69] Jonathan A. Parker and Christine Julliard, *Consumption Risk and the Cross Section of Expected Returns*, Journal of Political Economy, 2005, vol. 113, no. 1

[70] Ross, Stephen A. *The Arbitrage Theory of Capital Asset Pricing*, Journal of Economic Theory 13, 1976, 341 to 360.

[71] Steve Ross, *The Recovery Theorem*, Journal of Finance, Vol. LXX, No. 2, 2015

[72] Walter Rudin, *Functional Analysis*, Second Edition, McGraw-Hill Inc, 1991

[73] Federico Severino, *Isometric Operators on Hilbert spaces and the Classical Wold Decomposition for Stationary Time Series*, Boccini University Working Paper, September, 2014

[74] Daniel M. Stroock, *Introduction to Markov Processes*, Springer Graduate Texts in Mathematics, Vol. 230, Second Edition, 2014

[75] Terrance Tao and Van Vu, *Random Matrices: Universality of Local Eigenvalue Statistics up the the Edge*, UCLA Mathematics Working Paper, 2009

[76] S. R. S. Varadhan, *Large Deviations*, The Annals of Probability, Vol. 36, No. 2, 397-419, 2008

[77] Ivo Welch and Amit Goyal, *A Comprehensive Look at The Empirical Performance of Equity Premium Prediction*, Review of Financial Studies, v.21 n.4 2008