Submitted to *Unknown Journal*

A Machine Learning Methodology for Daily Assessment of Bank Health, Interconnectedness, and Systemic Risk

SHAWN MANKAD
North Carolina State University

CELSO BRUNETTI
Federal Reserve Board

JEFFREY H. HARRIS
American University

We propose a novel methodology to estimate the portfolio composition of banks as a function of daily stock returns. Building on a model where individual bank balance-sheets connect through common holdings, we derive and solve a constrained semi-non-negative matrix factorization problem where the rows (corresponding to banks) of one latent matrix factor (representing asset holdings) are subject to probability constraints. Although banks report assets at low frequencies, estimating our factorization over a rolling window allows us to derive daily estimates of bank portfolios. We validate our estimates of asset holdings by showing they closely match balance-sheet data reported in quarterly regulatory filings.

1. INTRODUCTION

Financial crises have accentuated the need for effective monitoring, oversight, and regulation of financial markets and institutions. This paper presents a new

This material is based upon work supported by the National Science Foundation under Grant No. 1633158 (Mankad). The views in this paper should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System. All errors and omissions, if any, are the authors' sole responsibility.

method to estimate common risk factors in the banking system from stock returns at a daily resolution, providing a timely and ongoing assessment of individual bank diversification and systemic risk. Specifically, we create a model of overlapping financial institution balance sheets to motivate a constrained-matrix factorization problem that is a special case of non-negative matrix factorization (NMF), where one factor is constrained to be non-negative, but a second factor can be composed of elements of any sign.[1] In extending the Semi-NMF of Ding et al. [2008], we subject the non-negative factor to probability constraints, which correspond to each bank's percentage holdings in different asset classes. We frame the problem with a Bayesian perspective, using the Dirichlet distribution as a prior to enforce the probability constraint, and use past bank-level disclosures to the Federal Deposit Insurance Corporation (FDIC) to calibrate the Dirichlet concentration parameters.

The factorization of stock returns produces daily estimates of individual bank asset portfolios, which we use to characterize risk within and among banks. Intuitively, we derive an index of portfolio concentration (bank-specific risk) for each individual bank, capturing exposure of a bank to asset-specific risk, and an index of portfolio similarity across banks, capturing the banking sector's vulnerability to propagating shocks (see e.g., Gai et al. [2011], Caccioli et al. [2014, 2015], Greenwood et al. [2015], Glasserman and Young [2015], Wang et al. [2019]). In this respect, our measures complement existing systemic risk measures that assume some stress scenario to capture losses of capital [Acharya et al., 2017, Brownlees and Engle, 2017], losses in asset values [Tobias and Brunnermeier, 2016], or losses due to fire sales for the banking sector when capital levels and assets are known [Shin and White, 2020, Duarte and Eisenbach, 2021]. We demonstrate the usefulness of our measures for prudential supervision and risk management by performing a detailed case study of the four banks that failed in the first quarter of 2023. We show that our indexes provide an early warning that these failed banks

[1] In this light, our problem differs from the standard Lee and Seung [1999] NMF, where the lower rank factors are both non-negative. This Semi-NMF allows our model to be applied in contexts where the input matrix is of mixed signs.

were insufficiently diversified or unusual in their asset holdings. By examining each bank's estimated portfolio holdings, we find that these early warning signals are interpretable and driven daily by real-world events.

We believe this work contains several contributions to the literature. Our work is the first to connect accounting models of balance sheets to matrix factorization techniques widely used in other domains. The form and estimation strategy of our factorization model is also new. Though several Bayesian NMF-style factorizations have been developed [Salakhutdinov and Mnih, 2008, Schmidt et al., 2009, Psorakis et al., 2011, Agarwal and Chen, 2010, Yang and Dunson, 2016], we are the first to provide a Bayesian formulation of the semi-NMF problem using Dirichlet priors to rigorously enforce probability constraints in an unsupervised setting. Our model and estimation approach are also the first to our knowledge to resolve the well-known issues of scale and rotational invariance with NMF-type factorizations [Paatero et al., 2002], resulting in estimates that are robust to the initialization – a key property for our setting in risk management and economic analysis. We also benchmark optimization versus Bayesian estimation strategies for constrained matrix factorization. [2] Lastly, we draw an important relationship between the proposed model and fuzzy K-means clustering [Bezdek et al., 1984] to shed light on which characteristics drive the model's favorable performance.

The rest of the paper is structured as follows. In §2 we examine the evolution of bank balance sheets to motivate our matrix factorization model. This derivation helps ground our model in economic principles and provides a clear interpretation to the factorization results. We also discuss in §2.2 how the factorization model can be used to measure concentration and similarity risk in the banking sector. In §3, we present our Bayesian formulation of the factorization model, including estimation details for all parameters. This is followed by validation of our

[2] Our results also indicate a trade-off between computing time and estimation quality. The heuristic of normalizing the derived solution ex post is fastest in run time but performs poorly in terms of accuracy for our application setting, while our Bayesian estimation is computationally intensive but consistently produces more stable and accurate solutions.

model and estimation using synthetic and real balance sheet data in §4. We conclude with an analysis of all publicly traded US banks in §5 and a short discussion in §6. Proofs and additional results are provided in the Appendix.

2. Model of Bank Balance Sheets

2.1 *A Stylized Accounting Model*

Our starting point for building the factorization model and subsequent risk measures is a basic accounting model where individual bank balance sheets connect through common holdings, aggregated to the industry level. Let there be $i = 1, \dots, n$ banks under consideration. N_{ikt} denotes the number of shares held in asset $k = 1, \dots, K$ (equities, bonds, commodities, etc.) on day t by bank i, and Y_{kt} denotes the market value of asset k on day t. Then $PV_{it} = \sum_k N_{ikt}Y_{kt}$ is the total market value of all bank i assets on day t. We can further define the percentage invested in each of the k assets by bank i on day t as $W_{ikt} = \frac{N_{ikt}Y_{kt}}{\sum_k N_{ikt}Y_{kt}}$, where $\sum_k W_{ikt} = 1$. Lastly, let E_{it} indicate the market value of bank i's equity on day t and let D_{it} be the total value of debt liabilities of bank i on day t. Note that non-negativity of W_{ikt} implies no short selling, which we believe is reasonable, given regulatory restrictions on bank portfolios and the intermediary role that banks play in the economy.

Consider a financial system in which banks connect lenders to borrowers as intermediaries, collecting deposits from households and firms and investing the deposits in a portfolio of assets, including loans to households (e.g. mortgages and consumer debt) and firms. The balance sheet for any individual bank i on day t can be partitioned as in Table 1.

Assets	Liabilities
$N_{i1t}Y_{1t}$	E_{it}
$\vdots$	
$N_{iKt}Y_{Kt}$	D_{it}

Table 1. Balance sheet representation for bank i.

Note that the balance-sheet model presented here slightly differs from previous literature [Shin, 2009, Elliott et al., 2014, Brunetti et al., 2019]. We omit the interbank market because this market dried up after the 2007-2009 crisis. Banks with excess reserves or in need of cash have since used FED Overnight Reverse Repurchase Agreements or repos/security agreements with other institutions. As investments, these operations are captured as assets in our model. In addition, while few institutions mark their balance sheets to market values, our approach uses the market value of equity as a proxy for the accounting values on the balance sheet.

The standard balance sheet identity, where assets equal liabilities, applied to Table 1 yields $\sum_k N_{ikt}Y_{kt} = E_{it} + D_{it}$. Taking first differences yields: $\sum_k \Delta(N_{ikt}Y_{kt}) = \Delta E_{it} + \Delta D_{it}$, which implies that

$$\Delta E_{it} = \sum_k \Delta(N_{ikt}Y_{kt}) - \Delta D_{it}. \tag{1}$$

Note that the left hand side is the one day ahead return on equity which we measure using stock returns. Recall that D_{it} represents debt claims on the the banking sector by households, mutual and pension funds, and other institutions. Following several previous works that utilize similar accounting models [Shin, 2009, Elliott et al., 2014, Brunetti et al., 2019], we assume that these debt liabilities evolve slowly, i.e., that $\Delta D_{it} = 0$. If we further assume that asset prices and bank-specific weights are stable within an appropriately short time interval, Proposition 1 establishes that the change in the market value of all bank i assets can be calculated using the weights W_{ikt} in place of the number of shares N_{ikt}.

PROPOSITION 1. *Assume that $\Delta W_{ikt} = \Delta Y_{kt} = 0$ for an appropriately small interval of time. Then $\Delta PV_{it} = \sum_k \Delta(W_{ikt}Y_{kt}) + \epsilon_{it}$, where ϵ_{it} is noise.*

To check our assumptions, we validate our model estimates using real balance data reported quarterly to the FDIC in §4.2. The results show that this assumption is reasonable in that our method produces accurate estimates of percentage holdings (W_{ikt}). Further, we find comparable results when validating the model with mutual funds in §4.3 where debt considerations are immaterial because funds are severely constrained to issue debt by law [Morley, 2013].

By Proposition 1 we can express Equation 1 as:

$$\Delta E_{it} = \sum_{k} \Delta(W_{ikt} Y_{kt}) + \epsilon_{it}. \tag{2}$$

Note that Equation 2 is not directly estimable because we observe the left-hand side (stock returns) with only n observations (1 per bank) for a given time point yet we want to infer the right-hand side which has more parameters ($(n+1)K$). To overcome this issue, we combine observations over a rolling window: Define $\boldsymbol{Z}_t = [\Delta \boldsymbol{E}_{t-T}, \Delta \boldsymbol{E}_{t-T+1}, \ldots, \Delta \boldsymbol{E}_t]$ (an $n \times T$ matrix). Also, using the assumption of weak stationarity of W_{ikt} in Proposition 1 (i.e., within the rolling window the expected value of W_{ikt} is fixed), in matrix notation the equation becomes:

$$\boldsymbol{Z}_t = \boldsymbol{W}_t \boldsymbol{V}_t + \boldsymbol{\epsilon}_t, \tag{3}$$

where $\boldsymbol{W}_t$ is an $n \times K$ matrix of percentages, $\boldsymbol{V}_t$ is an $K \times T$ matrix of real numbers (a function of asset returns), and $\boldsymbol{\epsilon}_t$ is an $n \times T$ noise matrix. Then, keeping with the spirit of matrix factorization as a lower dimensional embedding of the input data (i.e., $K \ll \min\{n, T\}$), the model is estimable and the portfolio composition is unique and identified (see §3 Proposition 3).

To establish a connection between the accounting and statistical perspectives of the proposed model, Proposition 2 states that the task of inferring percentage asset holdings for each bank can be viewed as a clustering problem:

PROPOSITION 2. *The proposed factorization in Equation 3 is a generalization of fuzzy K-means clustering [Jain, 2010] that allows for cluster-specific covariances, where the rank equals the number of clusters, the rows of $\boldsymbol{W}_t$ estimate the posterior probability of belonging to each cluster, and the columns of $\boldsymbol{V}_t$ capture the cluster's mean in T dimensional space.*

This relationship gives insight into when the proposed factorization will outperform fuzzy K-means, which can struggle when the data are not well approximated by Gaussian mixtures with identical variances. We see evidence of this in our real data as returns for small banks tend to have larger variance relative to large banks. In this case, our more general method is preferable because it can better match the data-generating process. For example, NMF and its variants have been extensively used for dimension reduction because they can achieve superior performance in many real-world settings [Lee and Seung, 1999, Ding et al., 2008, Li et al., 2021, Lassance et al., 2022].

2.2 *Measuring Concentration and Similarity Between Banks*

We use our factorization results to answer two key questions: Are bank portfolios well diversified? And how similar are portfolio holdings across banks? To answer the first question, we develop a concentration index that captures the degree of diversification of each bank's portfolio. To answer the second question, we develop a similarity index that captures how similar portfolio holdings are across banks. We view these two metrics as indicators of systemic vulnerabilities in the banking system. First, concentrated holdings on a small number of assets exposes a bank to asset-specific risk. *Ceteris paribus*, a bank with a portfolio concentrated only in one or two assets is more susceptible to shocks to those assets, increasing bank-specific risk.

Second, the similarity of asset holdings across banks suggests that shocks to any particular asset class will be borne across the entire banking system. The similarity of portfolio holdings is the theoretical justification of financial network analysis [Billio et al., 2012, Diebold and Yılmaz, 2014], and is based on a simple consideration: If two banks, A and B, hold the same asset and an exogenous shock forces A to liquidate the asset, the price of the asset will decline and therefore change the value of B's portfolio, potentially leading to B also selling the asset at an unfavorable price. Braverman and Minca [2018] describe how common asset holdings can transmit financial distress among banks. Wang et al. [2019] argue that portfolio similarity reflects similar banking business models and are therefore informative about systemic credit risk. Others establish that common asset holdings can amplify economic shocks, thereby raising the chances of simultaneous bank failures [Wagner, 2010, Beale et al., 2011, Gai et al., 2011, Caccioli et al., 2014, 2015, Greenwood et al., 2015]. In case of a shock to an asset class, our concentration measure captures the risk exposure of an individual bank whereas our similarity measure assesses the likelihood that the shock will propagate among banks.

An important advantage of our approach is to allow estimation of portfolio weights at a higher frequency than typically reported in official bank filings. In fact, we obtain daily estimates of $\boldsymbol{W}_t$ by employing daily stock returns as a proxy for the change in value of bank equity, $E_{it+1} - E_{it}$. After obtaining an estimate for $\boldsymbol{W}_t$, we calculate the average Herfindahl index for the banking sector of diversification/concentration across asset categories

$$\text{Concentration}_t = \frac{1}{n} \sum_{i,k} W_{ikt}^2. \tag{4}$$

To measure the similarity in assets held across banks, we define the average pairwise portfolio similarity between all pairs of banks i and j as

$$\text{Similarity}_t = \frac{1}{n^2} \sum_{i,j,k} \min(W_{ikt}, W_{jkt}). \tag{5}$$

The similarity index is bounded between 0 and 1, with zero values indicating that each pair of banks hold completely non-overlapping portfolios, whereas values equal to one indicate the banking sector holds identical portfolios.

3. A Bayesian Matrix Factorization Model

We denote the rows of a matrix $\boldsymbol{X}_t$ as $\boldsymbol{X}_{i.t}$ and columns as $\boldsymbol{X}_{.jt}$. Also $\boldsymbol{X}_{/x_{i.t}}$ denotes the matrix $\boldsymbol{X}_t$ excluding the i-th row.

The proposed factorization model in Equation 3 can be readily seen as a variant of the NMF problem, where $\boldsymbol{Z}_t$ is given and the objective is to estimate $\boldsymbol{W}_t$ and $\boldsymbol{V}_t$. Most works in NMF do not include the sum-to-one (STO) constraint for computational reasons, which contributes to a lack of identification. Specifically, estimates in NMF are always rescalable, where $\boldsymbol{W}_t$ can be multiplied by a positive constant c and $\boldsymbol{V}_t$ by $1/c$ to obtain different $\boldsymbol{W}_t, \boldsymbol{V}_t$ without changing their product. The recovered factors can also be rotated to produce different $\boldsymbol{W}_t, \boldsymbol{V}_t$ with the same product (i.e., $\boldsymbol{W}_t\boldsymbol{H}^T$ and $\boldsymbol{H}\boldsymbol{V}_t$ where $\boldsymbol{H}^T\boldsymbol{H} = \boldsymbol{I}$). Under conventional NMF formulations, it is not possible to differentiate a change in the percentage asset-class holding from the change in value of the given asset class, and further there are many optimal solutions. We resolve these issues by requiring the rows of $\boldsymbol{W}_t$ to sum-to-one (STO) and our estimation strategy enforces this constraint exactly.[3]

PROPOSITION 3. *Suppose $\boldsymbol{Z}_t = \boldsymbol{W}_t\boldsymbol{V}_t + \boldsymbol{\epsilon_t}$ such that the rows of $\boldsymbol{W}_t$ satisfy STO and non-negativity. Then if $\boldsymbol{W}_t\boldsymbol{H}^T$ and $\boldsymbol{H}\boldsymbol{V}_t$ are another valid solution, then $\boldsymbol{H}$ must be a permutation matrix.*

Because non-negative factors cannot be solved for analytically, the typical approaches in NMF pose an optimization problem based on minimizing an objective function like the Frobenius norm of the difference between $\boldsymbol{Z}_t$ and the estimated factors to obtain an estimate of $\boldsymbol{W}_t$ and $\boldsymbol{V}_t$ in Equation 3 [Berry et al., 2007, Lee and Seung, 1999]. When faced with STO constraints, the usual approach is to find approximate solutions (i.e., continuous relaxation of constraint using a La-Grangian penalty) or to ignore the constraint in the estimation and normalize the factors ex post in a second stage (see, e.g., Heinz and Chein-I-Chang [2001], Huck et al. [2010]). Both have computational advantages, but do not guarantee solutions that are stable to the estimation algorithm's random starting point. Moreover, due to the fundamental issues of rotational and scale invariance, conventional optimization methods can provide qualitatively different solutions, depending on the random seed, reducing the value of these methods for economic applications. To fully resolve these issues, we develop a novel Bayesian estimation framework that expresses the non-negativity and probability constraints using appropriate distributional assumptions with parameter estimation relying on Markov Chain Monte Carlo (MCMC) techniques.

[3] We acknowledge that in our proposed factorization model, the columns of $\boldsymbol{W}_t$ and correspondingly of $\boldsymbol{V}_t$ can be arbitrarily ordered. This is a common property of most factorization models other than the Singular Value Decomposition.

We assume that $\boldsymbol{Z}_t$ has the following conditional likelihood:

$$p(\boldsymbol{Z}_t|\boldsymbol{W}_t, \boldsymbol{V}_t, \sigma^2) = \prod_{i,t} \text{Normal}(\sum_k W_{ikt}V_{kt}, \sigma^2). \tag{6}$$

We use the Normal distribution for tractability and ease of computation, though this does not necessarily sacrifice the overall accuracy of the factorization. In fact, note that the variance of the Normal distribution is a random variable given by an Inverse Gamma density with shape η and scale θ

$$p(\sigma^2) = \text{Inverse Gamma}(\eta, \theta). \tag{7}$$

The Inverse Gamma distribution as a prior for σ^2 is a natural choice and extensively used in similar models [Korteweg and Sorensen, 2010], because for certain values of η and θ, the aggregate distribution of $\boldsymbol{Z}_t$ becomes heavy tailed and equivalent to the t-distribution which comports well with empirical distributions of stock returns [Upton and Shannon, 1979]. Specifically, in our empirical work we set the shape and rate parameters to 1, equivalent to a t-distribution with 10 degrees of freedom.

We also assume that each row of $\boldsymbol{W}_t$, denoted by $\boldsymbol{W}_{i.t}$, is distributed according to a Dirichlet distribution with the parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$.

$$p(\boldsymbol{W}_{i.t}) = \text{Dirichlet}(\boldsymbol{\alpha}) \tag{8}$$

The Dirichlet distribution, whose range is all discrete probability distributions of length K, is commonly used in nonparametric Bayesian statistics to model unknown probability distributions [Antoniak, 1974, Sethuraman, 1994]. α_k can take any value greater than zero. As α_k gets larger, probabilities for W_{ikt} are less concentrated and closer to uniform, meaning that the assets held in each bank portfolio and across banks are approximately equal. As α_k approaches zero, W_{ikt} is sparser (more weights are zero, though the zero components can vary among banks) and each bank's portfolio is more concentrated on a particular asset class.

Because $\boldsymbol{V}_t$ represents changes in asset values at the daily level, we expect its distribution to be unimodal and centered on a small constant, capturing market trends. We also expect the true distribution of $\boldsymbol{V}_t$ to have heavier tails, but we show that the Gaussian distribution offers a suitable approximation with computational advantages in that elements of $\boldsymbol{V}_t$ are assumed to be independently normally distributed with mean μ and variance σ_V^2:

$$p(\boldsymbol{V}_t) = \prod_{k,t} \text{Normal}(\mu, \sigma_V^2). \tag{9}$$

Proposition 4 states that although the prior distribution assumes daily returns between asset classes in $\boldsymbol{V}_t$ are independent, the correlation structure between asset returns is learned implicitly through the estimation, which we discuss below.

PROPOSITION 4. *The posterior distribution of $\boldsymbol{V}_t$ will exhibit correlations between asset classes.*

By Bayes rule, the joint posterior of $\boldsymbol{V}_t$ is proportional to

$$p(\boldsymbol{W}_t, \boldsymbol{V}_t, \sigma^2 | \boldsymbol{Z}_t) \propto p(\boldsymbol{Z}_t | \boldsymbol{W}_t, \boldsymbol{V}_t, \sigma^2) p(\boldsymbol{W}_t) p(\boldsymbol{V}_t) p(\sigma^2), \tag{10}$$

where we assume that $\boldsymbol{W}_t, \boldsymbol{V}_t$, and σ^2 are independently distributed, as in Equations 6 through 9. Computing the posterior densities $p(\boldsymbol{W}_t | \boldsymbol{Z}_t)$ and $p(\boldsymbol{V}_t | \boldsymbol{Z}_t)$ requires solving an intractable integral of the joint posterior distribution in Equation 10.[4] To overcome this challenge, we use a combination of standard MCMC methods. The basic idea is to construct a Markov chain that has the desired distribution as its limiting distribution. Thus, once the Markov chain has converged to its equilibrium, repeatedly sampling states of the chain provides an empirical estimate of the desired distribution that is accurate to an arbitrarily high degree. From this empirical distribution, the expectation can be readily calculated.

Because we can apply conjugate distributional properties to derive explicit closed forms of the posterior distributions for $\boldsymbol{V}_t$ and σ, conditional on the data $(\boldsymbol{Z}_t)$ and the current state of each of the parameters $(\boldsymbol{W}_t, \boldsymbol{V}_t, \sigma^2)$, we use Gibbs sampling to estimate the marginal distributions $p(\boldsymbol{V}_t | \boldsymbol{Z}_t)$ and $p(\sigma | \boldsymbol{Z}_t)$. In other words, the Markov chain is defined by the conditional posterior distributions

[4] Our estimation routines in a documented R package are available upon request.

and iterated until convergence, as in any MCMC method, after which samples are drawn and averaged to derive point estimates.

To estimate $p(\boldsymbol{W}_t|\boldsymbol{Z}_t)$, we use a more general version of Gibbs sampling, the Metropolis-Hastings algorithm, because the conditional posterior distribution of $\boldsymbol{W}_t$ is not composed of conjugate distributions and thus cannot be characterized analytically. The estimation procedure exploits the notion that we are still able to compute the value of a function that is proportional to the desired distribution (shown explicitly in the proof to Proposition 5). This proportion is used to generate Markovian samples iteratively that converge to the desired distribution as the number of samples grows.

PROPOSITION 5. *The posterior distributions for $\boldsymbol{V}_t$ can be empirically calculated via Gibbs sampling, where*

$$p(V_{kt}|\boldsymbol{Z}_t, \boldsymbol{W}_t, \boldsymbol{V}_{/v_{kt}}, \sigma^2) = \textit{Normal}(\mu_p, \sigma_p^2)$$
$$\sigma_p^2 = (\frac{\|\boldsymbol{W}_{.kt}\|_2^2}{\sigma^2} + \frac{1}{\sigma_V^2})^{-1}$$
$$\mu_p = \sigma_p^2(\frac{\tilde{\mu}\|\boldsymbol{W}_{.kt}\|_2^2}{\sigma^2} + \frac{\mu}{\sigma_V^2})$$
$$\tilde{\mu} = \frac{\boldsymbol{Z}_{.kt}^T\boldsymbol{W}_{.kt} - (\boldsymbol{V}_t^T\boldsymbol{W}_t^T)_{t.}\boldsymbol{W}_{.kt} + \|\boldsymbol{W}_{.kt}\|_2^2 V_{kt}}{\|\boldsymbol{W}_{.kt}\|_2^2}.$$

The posterior distributions for σ can be empirically calculated via Gibbs sampling, where

$$p(\sigma^2|\boldsymbol{W}_t, \boldsymbol{V}_t, \boldsymbol{Z}_t) = \textit{Inverse Gamma}(\eta', \theta')$$
$$\eta' = \eta + \frac{nT}{2} + 1$$
$$\theta' = \frac{1}{2}\sum_{i,t}(Z_{it} - \sum_k W_{ikt}V_{kt})^2 + \theta.$$

The posterior distribution for $\boldsymbol{W}_t$ can be empirically calculated using the Metropolis Hastings algorithms with a uniform proposal distribution. The candidate row $\widetilde{\boldsymbol{W}_{i.t}}$ is accepted with probability

$$\min\left(1, \frac{p(\widetilde{\boldsymbol{W}_{i.t}}|\boldsymbol{Z}_t, \boldsymbol{W}_{t/W_{i.}}, \boldsymbol{V}_t, \sigma^2)}{p(\boldsymbol{W}_{i.t}|\boldsymbol{Z}_t, \boldsymbol{W}_{t/W_{i.}}, \boldsymbol{V}_t, \sigma^2)}\right).$$

A fundamental question with MCMC methods is determining whether the Markov chain has converged. We utilize a convergence diagnostic proposed by Geweke [1992] based on a test of equality of the means of different portions of the Markov chain. The main idea is that if the samples are drawn from the stationary distribution of the chain, then the two means are equal and Geweke's statistic has an asymptotically standard normal distribution. Based on this diagnostic, we find in our real data analysis evidence of convergence around 10,000 iterations and thus base our estimates on 50,000 MCMC samples after 10,000 burn-in iterations.

Another important diagnostic for our MCMC is the acceptance rate of the Metropolis-Hastings, which speaks to whether an appropriate step size has been selected for the proposed distribution of $\boldsymbol{W}_{i.t}$. We select the step size by grid search and find that the acceptance rate on the banking data is 25.2%, which is near the asymptotically optimal rate of 23% [Robert and Casella, 1999].

4. Validation of the Model and Estimation

In this section, we validate our model and Bayesian estimation framework from a statistical perspective through simulation exercises and with real data by evaluating the estimates of $\boldsymbol{W}_t$ (the matrix of weights invested in each asset class) against actual balance-sheet data for banks and mutual funds.

4.1 *Simulation*

We compare our proposed model to techniques from the matrix-factorization and machine-learning literature that can be used to solve Equation 3:

1. The proposed factorization estimated with gradient descent techniques [Ding et al., 2008] and the STO constraint enforced ex post, i.e., the estimates

are normalized after each iteration, so that W adheres to probabilities (denoted as Semi-NMF with Normalization).

2. The proposed factorization with the STO constraint enforced via a Lagrangian penalty in the objective function (denoted as Semi-NMF with STO penalty). The penalty level is set to be a small constant (10^{-8}). The resultant estimates have row sums typically in $[0.7, 1.3]$. Estimates are normalized after estimation to satisfy sum-to-one constraints exactly.
3. The proposed factorization with Bayesian estimation (denoted as Bayesian Semi-NMF) with different parameters $\boldsymbol{\alpha} = \{0.1, 1\}$;
4. The Fuzzy K-means algorithm (denoted as FKmeans), which produces estimates of W based on a Gaussian mixture model [Bezdek et al., 1984];
5. Fuzzy analysis clustering (denoted as Fanny) with Euclidean distance as a measure of dissimilarity. We utilize the implementation in the "cluster" library of R in the function "fanny".

Driven by our application where $\boldsymbol{W}$ represents asset holdings and thus the distribution of elements in $\boldsymbol{W}$ has risk implications, we focus on this factor when assessing the performance of each method.[5] First, we assess the accuracy of the estimated $\boldsymbol{W}$ in terms of clustering accuracy. We report the adjusted Rand Index (ARI) using the nearest hard clustering of both the estimated and true $\boldsymbol{W}$ [Rand, 1971]. The ARI varies from zero to one, with larger values indicating more accurate estimates for $\boldsymbol{W}$. We also report the results of nonparametric hypothesis tests to compare the distribution of the true $\boldsymbol{W}$ with its estimate. The first distributional test is the Mann-Whitney U test [Mann and Whitney, 1947] to assess whether our estimate of $\boldsymbol{W}$ is stochastically smaller (or larger) than its true value.

The second, more stringent test we utilize is the Two Sample Anderson Darling Test, created by Scholz and Stephens [1987] based on the classical Anderson Darling Test [Anderson and Darling, 1954], to assess whether there are differences between the two samples with particular sensitivity at the tails of the sampled

[5] With some abuse of notation, we drop the time subscript in this subsection to improve readability.

distributions.[6] Note that element-wise accuracy comparisons for $\boldsymbol{W}$ or $\boldsymbol{V}$ (like mean squared errors) are not possible unless K is small, because the columns of each estimated factor can be ordered arbitrarily (a common property of factorization models).

We generate data using Equations 6 through 9. The number of columns of $\boldsymbol{Z}$ is fixed at $T = 30$ and the number of rows (firms) is set to be $n = 50$. We also set $(\mu, \sigma_V) = (0, 1)$ and the noise level to $(\eta, \theta) = (1, 1)$. The true and estimated ranks (K, number of underlying asset classes) are set equal to each other and varied between 2 and 10. We vary the Dirichlet parameter $\boldsymbol{\alpha} = \{0.1, 1\}$ to study how sparsity and concentration in $\boldsymbol{W}$ impacts estimation performance. When $\boldsymbol{\alpha} = 1$, the probabilities are closer to uniform (i.e. bank portfolios are more diversified across asset classes). This is a more challenging case from a clustering perspective since cluster membership overlaps heavily. When $\boldsymbol{\alpha} = 0.1$, the true bank portfolios (or cluster memberships) are more concentrated in a particular asset class.

Table 2 shows the detailed results averaged over all simulation instances. While the ARI shows that the Semi-NMF generally performs favorably relative to Fuzzy K-means and Fanny, there is a clear rank ordering within Semi-NMF methods depending on the estimation strategy. When the Dirichlet parameter is correctly specified, the Bayesian estimates consistently achieve the highest ARI, and is third best even when the Dirichlet parameter is badly misspecified. Further, when the parameter is correctly specified, the Bayesian estimation is the only technique to consistently produce estimates that pass both non-parametric hypothesis tests, i.e., the distribution of the estimated and true $\boldsymbol{W}$ are statistically indistinguishable.

To gain further insight into the role of estimation strategy, in Figure 1 we plot the empirical cumulative distribution function (ECDF) for the true synthetic $\boldsymbol{W}$ along with the ECDF of the estimated $\boldsymbol{W}$ from different random seeds. We see that normalizing the Semi-NMF produces estimates that converge to different

[6] While the Anderson Darling test is comparable to the Kolmogorov-Smirnov test, it has been shown in Monte-Carlo studies to have comparably greater statistical power [Razali et al., 2011]. With both tests, failing to reject the null hypothesis provides statistical evidence in favor of the validity of the model and estimation procedure.

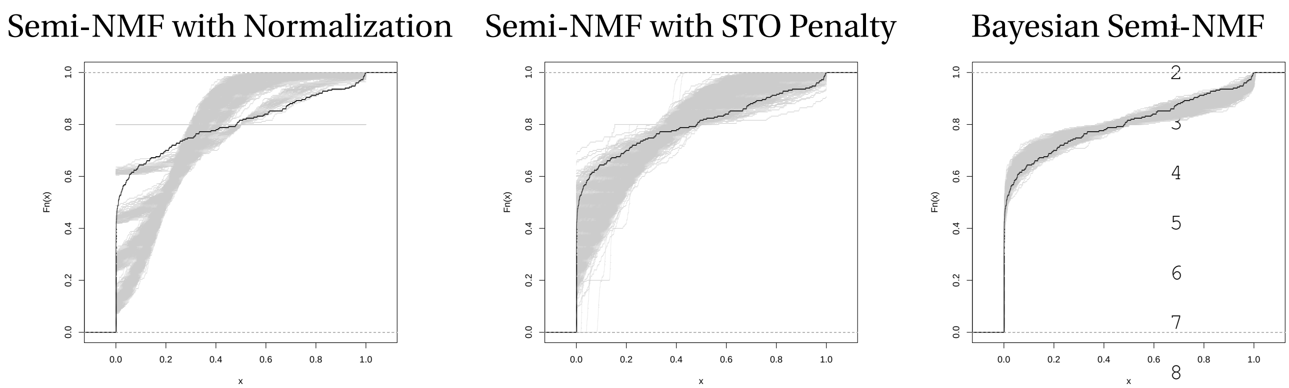

FIGURE 1. The solid black line shows the empirical cumulative distribution function (ECDF) for a single simulated instance of $\boldsymbol{W}$. Individual grey lines show the ECDF of each estimated $\boldsymbol{W}$ from 1000 different random seeds. The Bayesian estimation performs best in terms of stability and accuracy.

locally optimal points depending on the random seed, whereas the regularized Semi-NMF and Bayesian estimates more consistently converge around the true values of $\boldsymbol{W}$. The Bayesian Semi-NMF exhibits the least amount of noise and converges tightly on the true distribution. In summary, we show Bayesian estimates are more accurate and numerically stable compared to alternatives. In our application setting, the stability of the solution is an important property given that the ultimate goal is to help assess risk and inform policy makers.

4.2 *Validation with Banking Balance Sheet Data*

We consider all publicly traded banks insured by the Federal Deposit Insurance Corporation (FDIC). Specifically, we use data posted on the FDIC data repository (https://www.fdic.gov/bank/statistical/) of quarterly balance sheet variables that are reported by banks to the FDIC in regulatory filings. The data spans from 1998Q1 to 2020Q3 for 119 banks and includes the variables shown in Table 3. Note that these balance sheet items are straightforward to measure and used by

Scenario: Diversified Firm Portfolios ($\alpha = 1$)			
Method	ARI	MW	AD
Semi-NMF with Normalization	0.156	0.292	0.098
	(0.020)	(0.038)	(0.027)
Semi-NMF with STO Penalty	0.241	0.242	0.031
	(0.028)	(0.040)	(0.009)
Bayesian Semi-NMF ($\alpha = 0.1$)	0.235	0.188	0.003
	(0.028)	(0.036)	(0.001)
Bayesian Semi-NMF ($\alpha = 1$)	0.245	0.263	0.167
	(0.028)	(0.043)	(0.019)
FKmeans	0.175	0.150	0.000
	(0.023)	(0.038)	(0.000)
Fanny	0.196	0.150	0.000
	(0.025)	(0.038)	(0.000)
Scenario: Concentrated Firm Portfolios ($\alpha = 0.1$)			
Method	ARI	MW	AD
Semi-NMF with Normalization	0.456	0.125	0.000
	(0.031)	(0.037)	(0.000)
Semi-NMF with STO Penalty	0.682	0.241	0.000
	(0.028)	(0.036)	(0.002)
Bayesian Semi-NMF ($\alpha = 0.1$)	0.707	0.131	0.191
	(0.031)	(0.037)	(0.000)
Bayesian Semi-NMF ($\alpha = 1$)	0.636	0.125	0.000
	(0.030)	(0.037)	(0.000)
FKmeans	0.448	0.125	0.000
	(0.036)	(0.037)	(0.000)
Fanny	0.382	0.125	0.000
	(0.029)	(0.037)	(0.000)

TABLE 2. Simulation result averages over all ranks and trials with standard errors below in parentheses. For Mann-Whitney (MW) and Anderson Darling (AD) statistical tests, the average p-value is reported.

Abbreviation	Description	Mean	St. Dev
Cash	Cash and balances due from depository institutions	0.045	0.035
Securities	Total securities	0.202	0.091
Repo	Federal funds sold and reverse repurchase	0.016	0.028
Loans	Net loans and leases	0.663	0.101
Trade	Trading account assets	0.006	0.023
Bkprem	Bank premises and fixed assets	0.014	0.008
Ore	Other real estate owned	0.002	0.005
Intan	Goodwill and other intangibles	0.017	0.017
Idoa	All other assets	0.034	0.017

TABLE 3. Description of ground-truth bank balance sheet variables from the FDIC used in validation studies. Mean and standard deviation refer to the true $\boldsymbol{W}_t$ values over all banks and time points.

the FDIC to characterize the entire balance sheet of each bank. For each bank, we also collect their daily returns from the CRSP database.[7]

We first estimate the Bayesian Semi-NMF model using a 30-day rolling window. We set the number of factors $K = 9$ to match the number of balance sheet variables and set the parameter for the Dirichlet prior $\boldsymbol{\alpha}$ equal to the true average percentage holdings over all banks in the previous quarter in the FDIC data $(\frac{1}{n}\sum_i W_{ikt-1})$. We then compare the estimated and true W_{ikt} in two ways.

First, we validate our estimates by comparing the estimated and true concentration and similarity indexes. As shown in Figure 2, while the estimated indexes are noisier, they tend to exhibit the same local patterns and trends as seen in the true measures based on the FDIC quarterly filings. In fact, the average residual is -0.008 for the concentration index and -0.001 for the similarity index, demonstrating that our estimated indexes are also unbiased.

[7]For the CRSP database, we search under the following SIC codes: 6020 - Commercial banks, 6021 - National commercial banks, 6022 - State commercial banks, 6710 - Holding offices, 6712 - Offices of Bank Holding Companies, 6030 - Saving institutions.

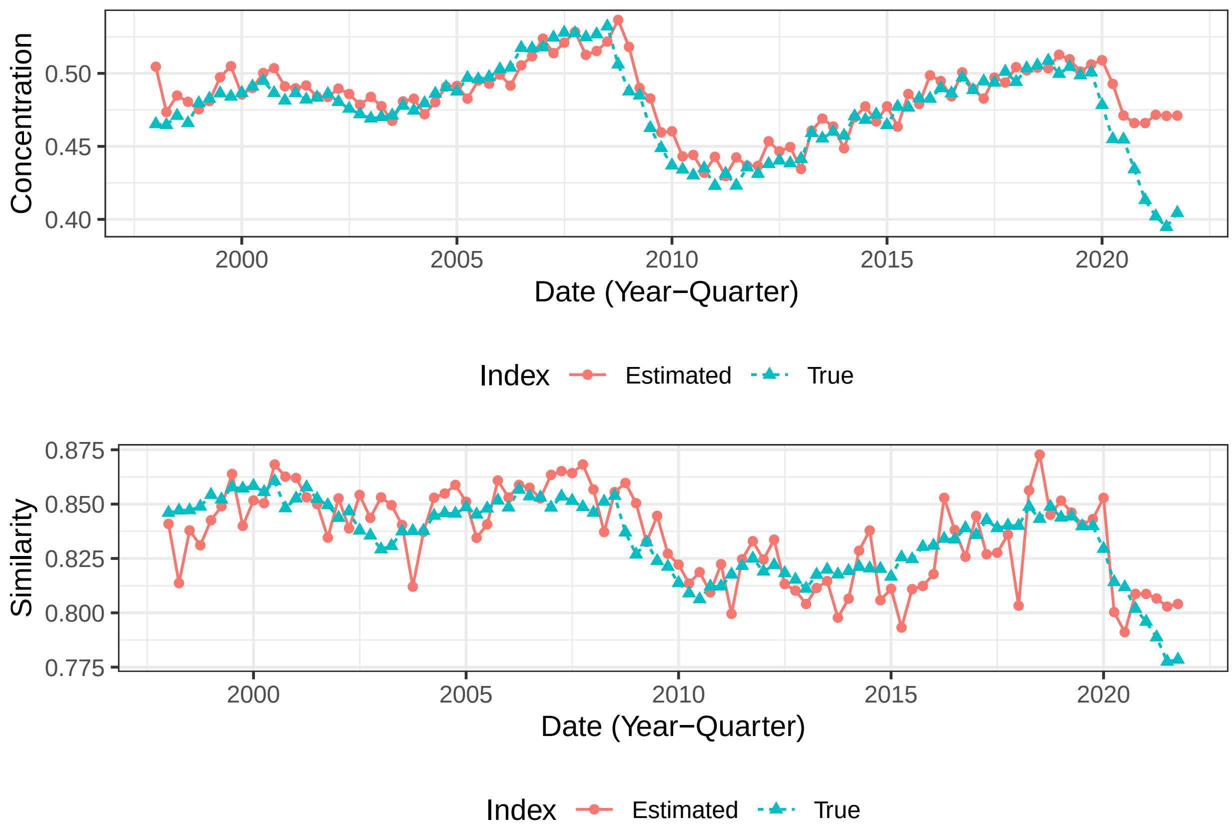

FIGURE 2. The true and estimated concentration and similarity indexes for publicly traded FDIC insured banks.

Second, we validate our estimates by calculating the average residual $\frac{1}{n}\sum_i(W_{ikt} - \hat{W}_{ikt})$ and mean-squared error $\frac{1}{n}\sum_i(W_{ikt} - \hat{W}_{ikt})^2$ for each quarter and balance sheet variable. The summary statistics in Table 4 show that the estimates from our factorization method are unbiased and accurate over the entire span of data: the average residual is close to zero and the average mean-squared error is no more than 0.011.

4.3 *Validation with Mutual Fund Balance Sheet Data*

While our main focus is on analyzing the banking sector, we note that since mutual funds are subject to strict reporting requirements and explicit about their

Variable	Statistic	Mean	St. Dev.	Min	Max
Cash	MSE	0.002	0.004	0.0001	0.017
	Residual	0.001	0.006	−0.018	0.024
Securities	MSE	0.007	0.004	0.001	0.018
	Residual	0.003	0.012	−0.025	0.047
Repo	MSE	0.001	0.002	0.0001	0.011
	Residual	−0.004	0.005	−0.020	0.013
Loans	MSE	0.011	0.006	0.001	0.029
	Residual	0.002	0.018	−0.064	0.040
Trade	MSE	0.0004	0.001	0.00004	0.010
	Residual	−0.005	0.002	−0.018	−0.0004
Bkprem	MSE	0.0005	0.002	0.00003	0.011
	Residual	0.001	0.002	−0.010	0.005
Ore	MSE	0.00004	0.0001	0.00001	0.001
	Residual	−0.002	0.001	−0.005	0.002
Intan	MSE	0.001	0.002	0.00003	0.009
	Residual	0.0002	0.004	−0.015	0.005
Idoa	MSE	0.001	0.003	0.0001	0.014
	Residual	0.001	0.004	−0.015	0.009

TABLE 4. Summary statistics over all quarters of the average mean-squared error and average residual for bank-level balance-sheet item estimates.

investment strategy with respect to different asset classes, this setting serves as a good alternative test bed for our methodology.

We consider all Refinitiv eMaxx funds that invested at least 10% into four or more market sectors from among the twelve shown in Table 5. The twelve market sectors characterize essentially all types of mutual funds, spanning bonds, stocks, real-estate, and other securities. For example, asset-backed securities

Abbreviation	Description	Mean	St. Dev
Com	Common stocks	0.149	0.205
Pref	Preferred stocks	0.006	0.026
Conv	Convertible bonds	0.008	0.043
Corp	Corporate bonds	0.261	0.143
Muni	Municipal bonds	0.013	0.022
Govt	Government bonds	0.223	0.140
Cash	Cash	0.011	0.161
ABS	Asset backed securities	0.094	0.115
MBS	Mortgage backed securities	0.148	0.139
EqOther	Equities other than common and preferred stocks	0.010	0.020
FixOther	Other fixed income securities	0.039	0.082
Oth	Other securities	0.038	0.150

TABLE 5. Description of ground-truth mutual fund balance sheet variables from the Refinitiv eMaxx database used in validation studies. Mean and standard deviation refer to the true $\boldsymbol{W}_t$ values over all funds and time points.

and mortgage-backed securities help characterize mutual funds that hold assets relating to mortgages. Unlike mortgage-backed assets, asset-backed are higher risk (lower FICO scores, omitted documentation, etc.) and do not quality for Government-Sponsored Enterprises (e.g., Fannie Mae and Freddie Mac). We also collect daily returns for each of the 121 funds to estimate our factorization. The data spans from 2010Q2 to 2022Q3.

We estimate the Bayesian Semi-NMF model using a 90-day rolling window. We set the number of factors $K = 12$ to match the number of market sectors and set the parameter for the Dirichlet prior $\boldsymbol{\alpha}$ equal to the true average percentage holdings over all funds in the previous quarter ($\frac{1}{n}\sum_i W_{ikt-1}$). We follow the same validation strategy as with the banking data by comparing (i) the estimated and true concentration and similarity indexes and (ii) the estimated and true percentage holdings.

Figure 3 shows that the estimated concentration index is persistently lower than the true value by 0.012 on average and the similarity index tends to be

FIGURE 3. The true and estimated concentration and similarity indexes for mutual funds.

slightly higher by 0.008 on average. These relatively small biases are caused by sparsity in the data. Roughly 27% of entries in the true W_{ikt} are equal to zero, and the estimation procedure will assign small but non-zero percentages for all of such entries which affects the overall concentration and similarity scores. Nonetheless, the estimated indexes exhibit the same trends and local patterns as seen in the true measures.

Table 6 presents summary statistics for the average residual and mean-squared error of our estimates. As with the banking data results, our factorization method produces estimates that are unbiased and accurate over entire span of data: the average residual is close to zero and the average mean-squared error is no more than 0.001.

Variable	Statistic	Mean	St. Dev.	Min	Max
Com	MSE	0.001	0.001	0.0002	0.004
	Residual	−0.004	0.004	−0.016	0.004
Pref	MSE	0.0001	0.0001	0.00004	0.001
	Residual	−0.003	0.001	−0.004	0.001
Conv	MSE	0.0002	0.0003	0.00004	0.002
	Residual	−0.005	0.002	−0.012	−0.0002
Corp	MSE	0.002	0.001	0.001	0.005
	Residual	0.006	0.009	−0.019	0.045
Muni	MSE	0.0002	0.0001	0.0001	0.001
	Residual	−0.00000	0.002	−0.005	0.005
Govt	MSE	0.002	0.001	0.001	0.009
	Residual	0.006	0.009	−0.024	0.026
Cash	MSE	0.003	0.002	0.0004	0.008
	Residual	−0.002	0.006	−0.017	0.016
ABS	MSE	0.001	0.0003	0.0002	0.001
	Residual	0.003	0.003	−0.008	0.010
MBS	MSE	0.002	0.002	0.0004	0.009
	Residual	0.004	0.006	−0.009	0.017
EqOth	MSE	0.0002	0.0003	0.0001	0.002
	Residual	−0.003	0.002	−0.010	0.004
FixOth	MSE	0.002	0.002	0.0001	0.013
	Residual	−0.001	0.006	−0.013	0.015
Oth	MSE	0.002	0.001	0.0005	0.008
	Residual	−0.002	0.007	−0.015	0.020

TABLE 6. Summary statistics over all quarters of the average mean-squared error and average residual for mutual fund-level balance-sheet item estimates.

Overall, we find that results for the mutual fund data are consistent with the banking data validation results. Our method produces accurate estimates for balance sheet items as well as the concentration and similarity indexes.

5. Analyzing the U.S. Banking Sector

We obtain daily stock returns from January 1, 1990 through April 28, 2023 for all U.S. publicly traded banks in the CRSP database. By considering all publicly traded banks, our sample size increases to 994 banks compared to our validation study of the 129 publicly traded banks that are FDIC insured. While obtaining and organizing ground-truth balance sheet information for so many banks can be a non-trivial process, our method uses only stock returns and can be estimated at a higher resolution than with regulatory disclosures. We categorize each bank in our sample into three size tiers, small (large) banks have median market capitalization in the lowest (highest) 25% and medium sized banks fall within the middle 50% of market capitalization among all banks. We provide summary statistics for our sample banks in Table 7, noting that smaller banks experience more daily return volatility.

Sample	n	Mean	St. Dev.
Large	177	0.000015	0.0291
Medium	518	-0.000164	0.0326
Small	303	0.000209	0.0495

TABLE 7. Summary statistics of daily stock returns for all publicly traded banks.

We estimate the Bayesian Semi-NMF model on this data using a 30-day rolling window. As in our validation study, set the number of factors $K = 9$ to match the number of balance sheet variables and set α equal to the true average percentage holdings over all banks in the previous quarter in the FDIC data ($\frac{1}{n}\sum_i W_{ikt-1}$). Prior to 1998 when FDIC data is not available, we set α equal to the true average percentage holdings over all banks and quarters in the FDIC data ($\frac{1}{nT}\sum_{i,t} W_{ikt}$).

5.1 *Balance Sheet Holdings, Concentration, and Similarity*

Figure 4 shows the estimated percentage holdings for different balance sheet items for the overall banking sector and each of the size tiers. As shown, the balance sheet variables for the banking sector were relatively stable in the 1990's compared to more recent time periods. Starting in the 2000's, we see variation in several balance sheet items including the two largest variables, securities and net loans. In fact, these two variables were at their maximum (net loans and leases) and minimum (total securities) levels heading into the 2007-2009 housing crisis. Cash holdings began to rise immediately following the passage of the Troubled Assets Relief Program. In response to COVID-19, net loans and leases decreased for the entire banking sector, while cash holdings sharply increased.

As shown in Table 8, small banks have lower levels of similarity and are also more concentrated. This empirical finding matches the real-world structure of the banking sector: Driven in part by the Community Reinvestment Act, a federal law that gives small banks incentives to invest in local municipal bonds, local companies, and real estate, small banks tend to be heavily focused in the areas they operate geographically and as a result should have a higher concentration and lower similarity with the rest of the sector than mid-sized and larger banks.

Sample	Concentration	Similarity
Large	0.396 (0.033)	0.904 (0.032)
Medium	0.409 (0.035)	0.901 (0.027)
Small	0.415 (0.045)	0.877 (0.036)

TABLE 8. Mean (standard deviation) of the estimated concentration and similarity for all publicly traded banks.

Figure 5 displays the average Herfindahl index of bank-asset concentration over time for each size tier. Consistent with the detailed estimates in Figure 4, concentration in the banking sector was relatively stable in the 1990s but began

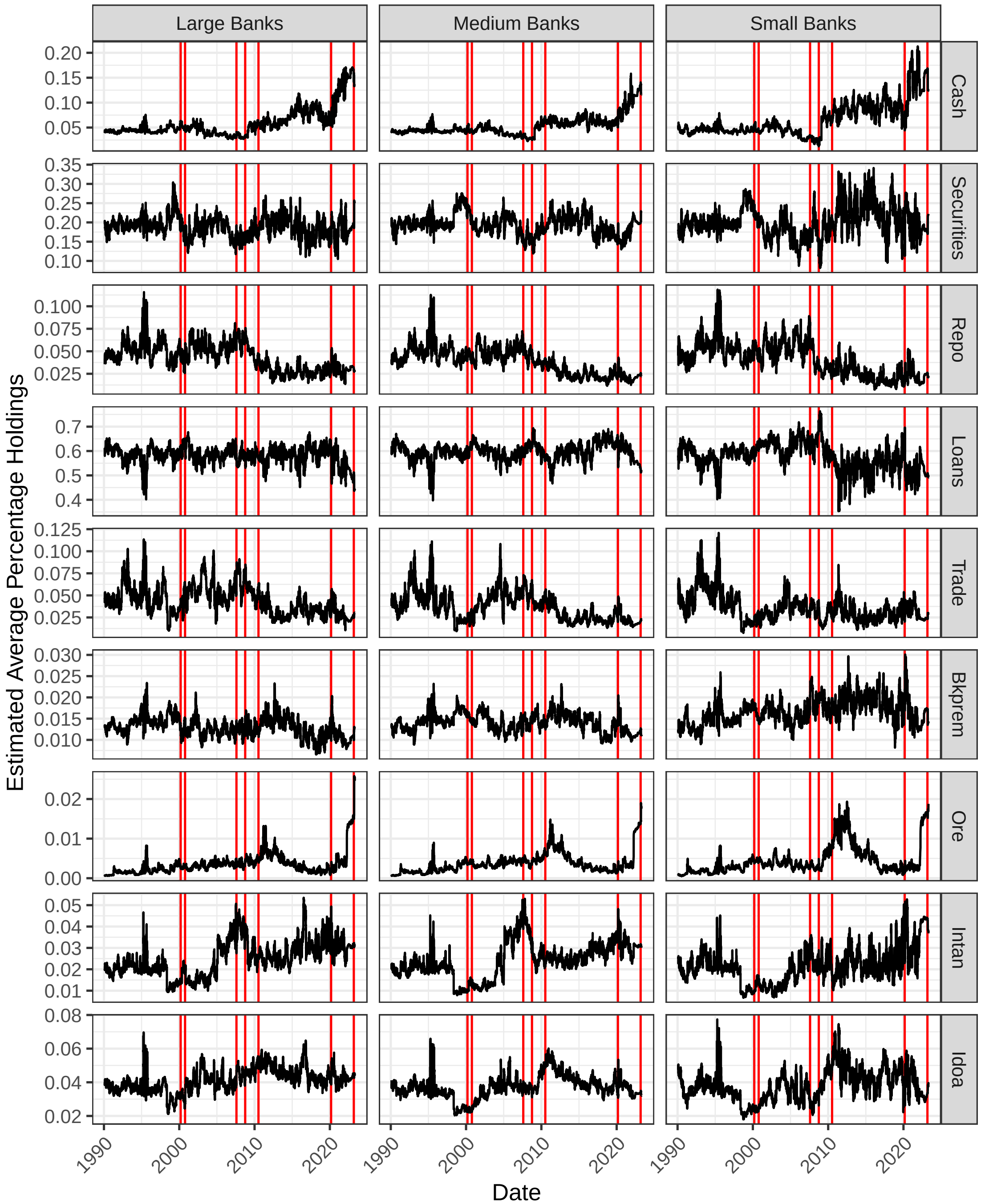

FIGURE 4. Estimated balance sheet item holdings over time. The vertical lines denote events: (1) March 10, 2000 and (2) October 9, 2000 corresponding to the peak and trough of NASDAQ during the dotcom bubble; (3) August 7, 2007 when the 2007-2009 housing crisis began; (4) October 3, 2008 when the Troubled Assets Relief Program was passed; (5) July 10, 2010 when The Dodd-Frank Wall Street Reform and Consumer Protection Act was passed; (6) March 1, 2020, onset of COVID-19 in the United States; (7) March 8, 2023 marking the first post-COVID bank failure.

to rise (due to the increase in net loans and leases holdings) in the mid-2000s until it reached its peak with the 2007-2009 housing crisis. Notably, concentration throughout the banking sector rose with passage of the Dodd-Frank Wall Street Reform and Consumer Protection Act, which coincided with increased cash holdings. As noted above, the onset of COVID-19 also led to increased cash holdings, but also to substantial decreases in the two largest balance sheet items: net loans and leases and total securities. As such, the concentration index decreased slightly following the onset of COVID-19.

Figure 6 displays the average of the similarity of a bank's assets to the rest of the banking sector over each size tier. As shown, we see consistently high similarity levels from 1990 until the housing crisis. For large banks especially, similarity decreased during this time until after the passage of the Troubled Assets Relief Program when similarity started to revert to pre-crisis levels. The volatility of similarity increased following the 2010 passage of the Dodd-Frank Wall Street Reform and Consumer Protection Act through COVID-19, which created a negative shock. By 2023, similarity had reverted to its pre-pandemic levels.

5.2 *2023 US Bank Crisis*

The 2023 failures of Silvergate Bank (SI), Silicon Valley Bank (SVB), Signature Bank (SBNY), and First Republic Bank (FRC) have re-focused attention on the health and stability of the US banking sector. We show next that our indexes provide an effective early warning that the failed banks were insufficiently diversified or unusual in their asset holdings. By examining each bank's estimated percentage holdings, we find that these early warning signals are interpretable, meaningful, and driven by real-world events.

Figure 7 shows the estimated concentration and similarity from December 15, 2022 to April 28, 2023 for the four failed banks with the 99% confidence interval of the average over all medium-sized banks. The four failed banks had much lower similarity and (with the exception of First Republic Bank) were also far more concentrated in their asset holdings, falling well outside the 99% confidence interval.

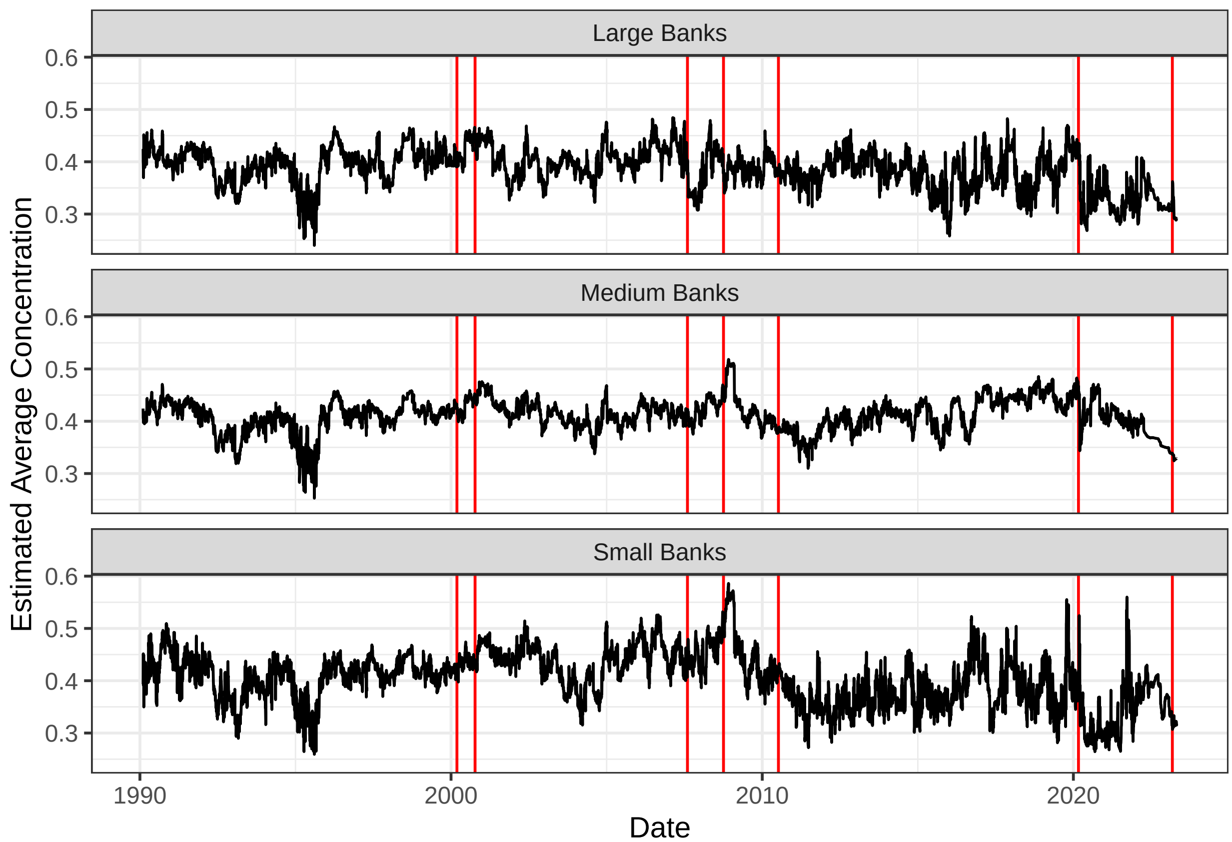

FIGURE 5. Concentration index over time. The vertical lines denote the same events: (1) March 10, 2000 and (2) October 9, 2000 corresponding to the peak and trough of the NASDAQ during the dotcom bubble; (3) August 7, 2007 when the 2007-2009 housing crisis began; (4) October 3, 2008 when the Troubled Assets Relief Program was passed; (5) July 10, 2010 when The Dodd-Frank Wall Street Reform and Consumer Protection Act was passed; (6) March 1, 2020, onset of COVID-19 in the United States; (7) March 8, 2023 marking the first bank failure in the post-COVID era.

Silvergate announced plans to liquidate and cease operations on March 8, Silicon Valley Bank and Signature Bank failed on March 10, and First Republic Bank would be acquired by JP Morgan on May 1. Leading up to its failure, Silvergate in particular was consistently the most dissimilar bank among all medium-sized banks, which comports with their uniquely high exposure to the cryptocurrency industry. From March 27 onward, First Republic Bank had a sudden drop in its

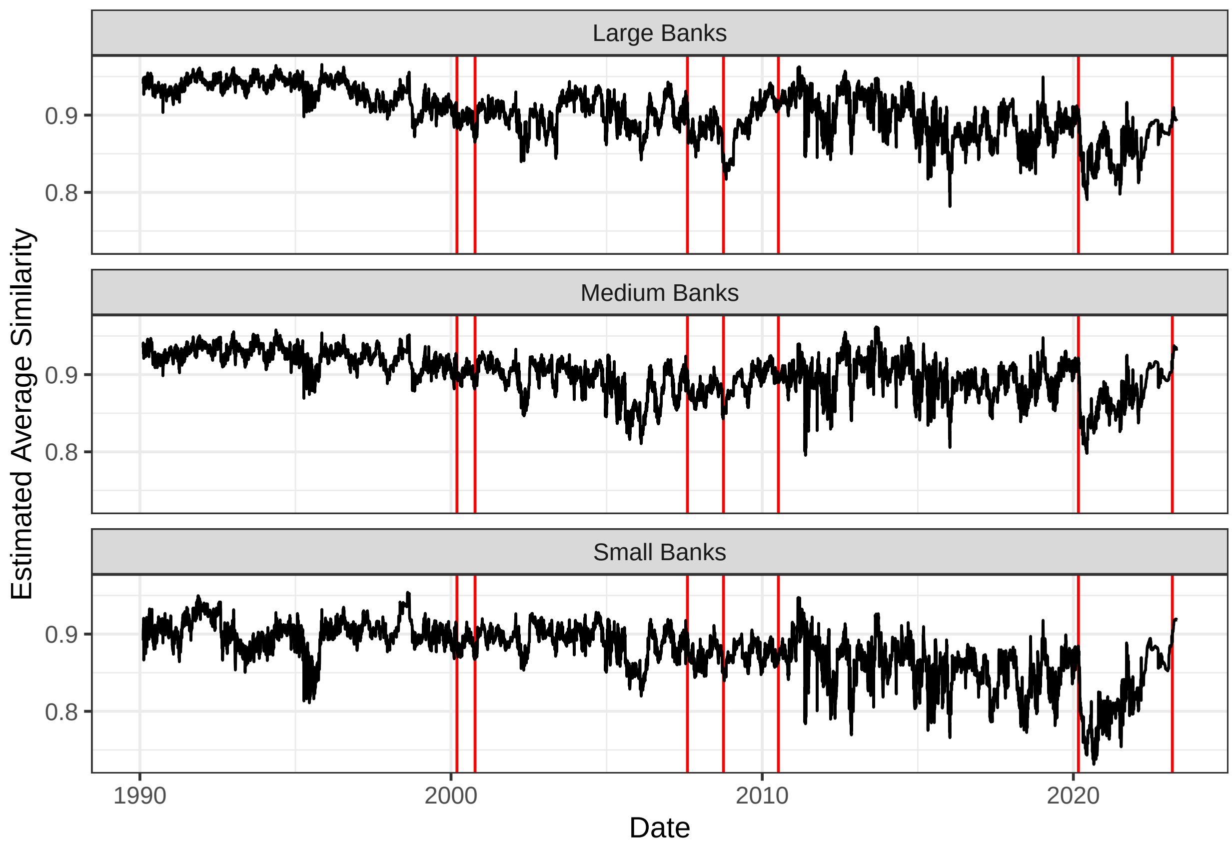

FIGURE 6. Similarity index over time. The vertical lines denote same events: (1) March 10, 2000 and (2) October 9, 2000 corresponding to the peak and trough of the NASDAQ during the dotcom bubble; (3) August 7, 2007 when the 2007-2009 housing crisis began; (4) October 3, 2008 when the Troubled Assets Relief Program was passed; (5) July 10, 2010 when The Dodd-Frank Wall Street Reform and Consumer Protection Act was passed; (6) March 1, 2020, onset of COVID-19 in the United States; (7) March 8, 2023 marking the first bank failure in the post-COVID era.

estimated similarity, making it the most dissimilar bank until it was acquired five weeks later.

For a more granular view of these banks' assets, Figure 8 presents estimated percentage holdings of key balance sheet variables. Focusing first on Silvergate Bank, Silicon Valley Bank, and Signature Bank, several notable findings emerge: (i) these banks held much lower levels of cash compared to the average bank; (ii)

Silvergate's assets were highly concentrated in securities; (iii) Silicon Valley Bank and Signature Bank were more concentrated in their loan portfolios.

For First Republic Bank, our method detects volatility in cash holdings, a sharp decrease in securities and loans, and an increase in reverse repurchase (repos) holdings from mid-March onward. These movements reflect real-world activity: The first spike in cash holdings corresponds to a March 16 rescue attempt by eleven large American banks depositing $30 billion with First Republic. However, high-net-worth customers (whose assets exceeded FDIC protection limits) continued to withdraw funds, drawing cash down, a result we uncover in late March that was only formally confirmed later in April. The sudden changes in securities and loans in late March correspond to growing concerns about the bank's balance sheet. First Republic's market value continued to drop precipitously throughout March and its credit rating was downgraded by S&P on March 19, reflecting the outflow from deposits and degradation of the bank's loan portfolio due to rising interest rates. Further, the majority of the bank's long term assets were in municipal bonds which are not eligible collateral for emergency Federal Reserve loans, so First Republic increasingly relied on reverse repos to raise funds as our estimated increase in Repo holdings suggests.

To further illustrate how our model can be useful for prudential supervision and risk management, we rank banks that exhibit tail behavior in their estimated index and cash holdings using the following metric:

$$\begin{aligned}\text{Outlier Score}_{it} = &(\text{Concentration}_{it} - UB^{Concentration}_{0.99,t}) \\ &+(LB^{Similarity}_{0.99,t} - \text{Similarity}_{it}) \\ &+(LB^{Cash}_{0.99,t} - \text{Cash}_{it}),\end{aligned} \tag{11}$$

where $UB^{X}_{0.99,t}$ and $LB^{X}_{0.99,t}$ are the upper and lower bound, respectively, for the 99% confidence interval for variable X. Note that each component of the outlier

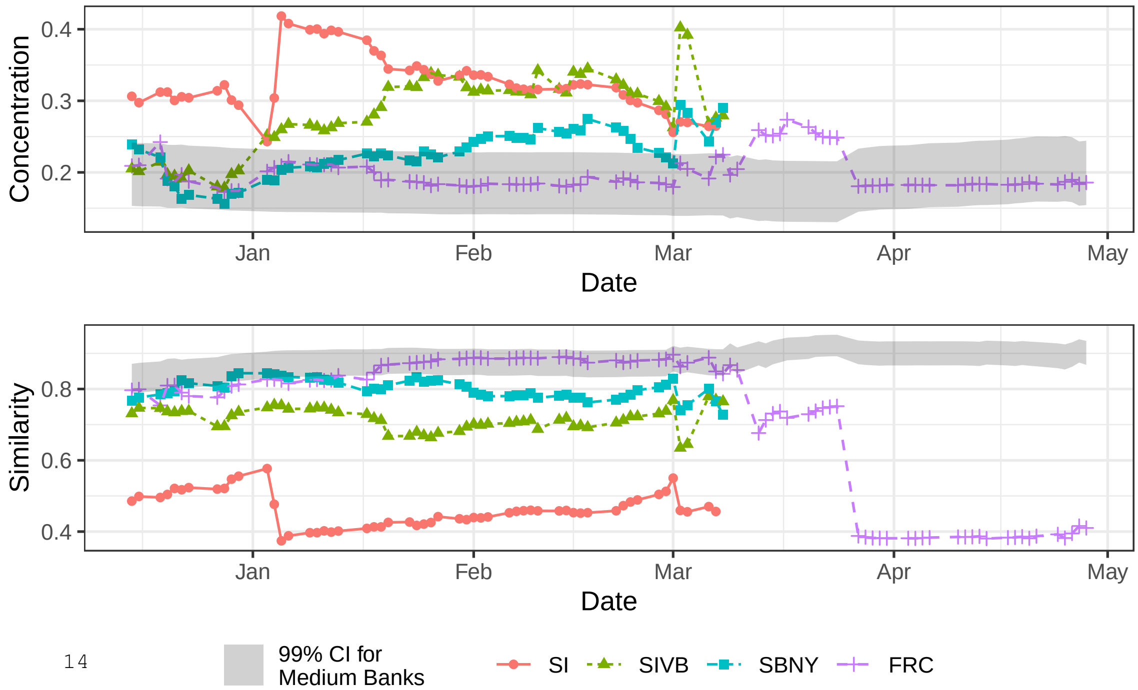

FIGURE 7. Estimated daily Concentration and Similarity measures from December 15, 2022 to April 28, 2023.

score represents a possible risk: (i) High concentration levels can indicate exposure to asset-specific risk; (ii) Low similarity with the banking sector is a risk indicator when the banking sector is generally healthy such as post Dodd-Frank;[8] (iii) low cash holdings are associated with several financial vulnerabilities.

Panel A of Table 9 shows that Silvergate Bank, Silicon Valley Bank, and Signature Bank stand out before their respective collapses. In fact, our methods consistently identify these three banks as problematic weeks before they collapsed. Of course, other banks have also had their credit rating downgraded and/or suffered major losses in market value during this time frame, including most of the bottom ten banks exhibiting tail behavior (see Table 9 Panel B from March 11, 2023 and Panel C from April 28, 2023).

[8] This is largely due to the fact that the FDIC asset categories are coarse. Impulse response functions in Appendix B show that higher similarity is associated with lower future levels of SRISK.

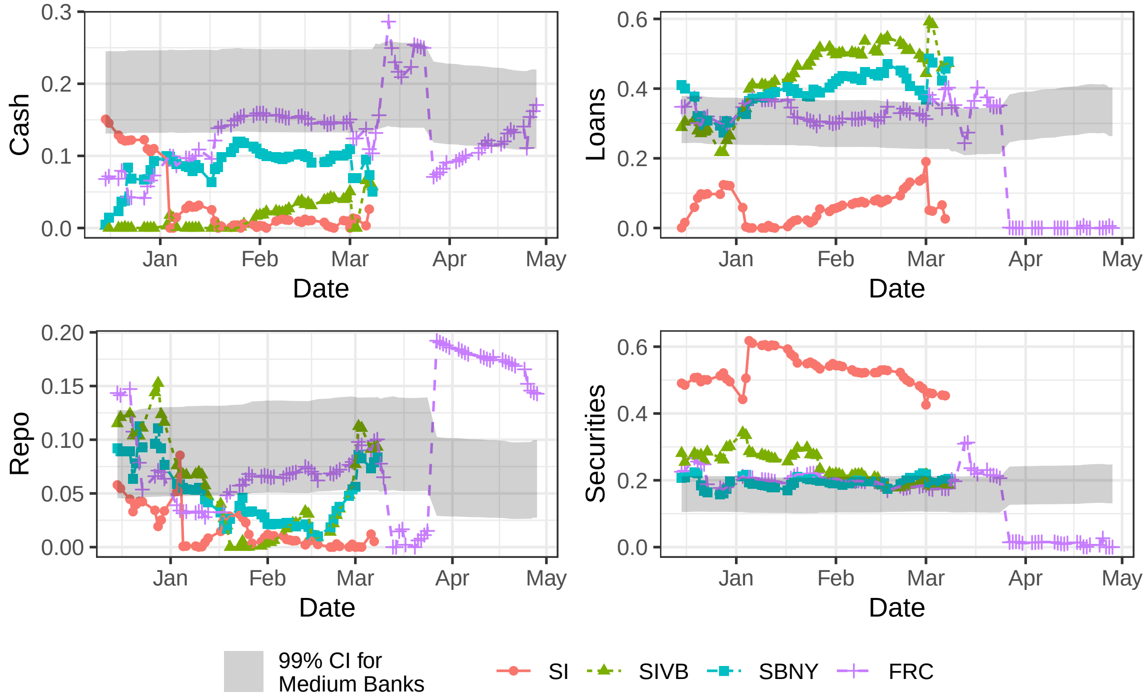

FIGURE 8. Estimated daily balance sheet percentage holdings from December 15, 2022 to April 28, 2023.

6. CONCLUSION

In this paper, we propose a novel method based on semi-non-negative matrix factorization to estimate the portfolio of bank assets as a function of daily stock returns. Using our estimates of bank holdings, we construct two daily risk measures: individual-bank portfolio concentration and common holdings across banks. We find evidence that these risk measures lead, in a forecasting sense, SRISK. We also identify banks as troubled well before their actual failure in the first quarter of 2023, suggesting that our method can be a useful addition to risk management tools [Engle and Ruan, 2019, Glasserman and Young, 2016, Cont and Schaanning, 2019, Zenios et al., 2021] to assess and forecast systemic risk in a timely manner. From an economic standpoint, our work solves a key issue for risk analysis by providing meaningful and timely information about individual bank holdings, interconnectedness, and systemic risk in the banking system.

Indeed, we demonstrate that our methods can generate distributions of bank asset holdings that can be utilized to identify potential problematic banks in advance of required regulatory filings. On February 15, 2023, for instance, we identify Silvergate Bank, Silicon Valley Bank, and Signature Bank as the largest outliers in estimated concentration, similarity and cash holdings–these banks subsequently failed on March 8, 10, and 10, respectively.

From a modeling standpoint, we advance the work on semi-non-negative matrix factorization to include a Bayesian component with a sum-to-one constraint. Motivated by an accounting model of bank balance sheets, we subject the rows of the non-negative factor (W) to a strict sum-to-one constraint and show that the strict enforcement of this constraint via a Bayesian formulation outperforms alternative optimization-based algorithms from the NMF and clustering literature. Our model also addresses scale and rotational invariance by the sum-to-one constraint. Our validation experiments show that our proposed approach produces solutions that are stable, accurate, and closely match holdings reported in regulatory filings.

An important area of future work is on improving the scalability of the estimation algorithm. Specifically, while we find that the MCMC approach has several important theoretical advantages, its computational cost can be prohibitively expensive with a large number of asset classes using a large rolling window length. In fact, in text analysis, for example MCMC techniques have generally lost popularity due to the rise of variational inference techniques that tend to be faster and easier to scale (though not as theoretically sound). A thorough comparison of variational versus our MCMC techniques may lead to important improvements for estimation algorithms. Lastly, while we focus mainly on the banking sector, we believe our results for mutual funds can be expanded to identify potentially important systemic risk/contagion driven by large mutual funds.

left=2cm

References

Viral V Acharya, Lasse H Pedersen, Thomas Philippon, and Matthew Richardson. Measuring systemic risk. *The Review of Financial Studies*, 30(1):2–47, 2017. [2]

Deepak Agarwal and Bee-Chung Chen. flda: matrix factorization through latent dirichlet allocation. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 91–100, 2010. [3]

Theodore W Anderson and Donald A Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769, 1954. [14]

Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174, 1974. [10]

Nicholas Beale, David G Rand, Heather Battey, Karen Croxson, Robert M May, and Martin A Nowak. Individual versus systemic risk and the regulator's dilemma. *Proceedings of the National Academy of Sciences*, 108(31):12647–12652, 2011. [7]

Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007. [9]

James C Bezdek, Robert Ehrlich, and William Full. Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203, 1984. [3, 14, 41]

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 1997. [41]

Monica Billio, Mila Getmansky, Andrew W Lo, and Loriana Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 104(3):535–559, 2012. [7]

Anton Braverman and Andreea Minca. Networks of common asset holdings: aggregation and measures of vulnerability. *The Journal of Network Theory in Finance*, 4(3), 2018. [7]

Christian Brownlees and Robert F Engle. Srisk: A conditional capital shortfall measure of systemic risk. *The Review of Financial Studies*, 30(1):48–79, 2017. [2]

Celso Brunetti, Jeffrey H Harris, Shawn Mankad, and George Michailidis. Interconnectedness in the interbank market. *Journal of Financial Economics*, 133(2): 520–538, 2019. [5]

Fabio Caccioli, Munik Shrestha, Cristopher Moore, and J Doyne Farmer. Stability analysis of financial contagion due to overlapping portfolios. *Journal of Banking and Finance*, 46:233–245, 2014. [2, 7]

Fabio Caccioli, J Doyne Farmer, Nick Foti, and Daniel Rockmore. Overlapping portfolios, contagion, and financial stability. *Journal of Economic Dynamics and Control*, 51:50–63, 2015. [2, 8]

Stanley S Chang and Chi-Kwong Li. Certain isometries on rn. *Linear Algebra and its Applications*, 165:251–265, 1992. [41]

Rama Cont and Eric Schaanning. Monitoring indirect contagion. *Journal of Banking and Finance*, 104:85–102, 2019. [32]

Francis X Diebold and Kamil Yılmaz. On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, 182(1):119–134, 2014. [7]

Chris HQ Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55, 2008. [2, 7, 13]

Fernando Duarte and Thomas M Eisenbach. Fire-sale spillovers and systemic risk. *The Journal of Finance*, 76(3):1251–1294, 2021. [2]

Matthew Elliott, Benjamin Golub, and Matthew O Jackson. Financial networks and contagion. *American Economic Review*, 104(10):3115–53, 2014. [5]

Robert F Engle and Tianyue Ruan. Measuring the probability of a financial crisis. *Proceedings of the National Academy of Sciences*, 116(37):18341–18346, 2019. [32]

Prasanna Gai, Andrew Haldane, and Sujit Kapadia. Complexity, concentration and contagion. *Journal of Monetary Economics*, 58(5):453–470, 2011. [2, 7]

John Geweke. Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian Statistics*, 4:641–649, 1992. [13]

Paul Glasserman and H Peyton Young. How likely is contagion in financial networks? *Journal of Banking and Finance*, 50:383–399, 2015. [2]

Paul Glasserman and H Peyton Young. Contagion in financial networks. *Journal of Economic Literature*, 54(3):779–831, 2016. [32]

Robin Greenwood, Augustin Landier, and David Thesmar. Vulnerable banks. *Journal of Financial Economics*, 115(3):471–485, 2015. [2, 8]

D.C. Heinz and Chein-I-Chang. Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 39(3):529–545, 2001. [9]

Alexis Huck, Mireille Guillaume, and Jacques Blanc-Talon. Minimum dispersion constrained nonnegative matrix factorization to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 48(6):2590–2602, 2010. [9]

Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010. [6]

Nathaniel Johnston. Isometries of the vector p-norms: Signed and complex permutation matrices. *Blog*, 2010. URL http://www.njohnston.ca/2010/09/isometries-of-the-vector-p-norms-signed-and-complex-permutation-matrices/. [41]

Arthur Korteweg and Morten Sorensen. Risk and return characteristics of venture capital-backed entrepreneurial companies. *The Review of Financial Studies*, 23(10):3738–3772, 2010. [10]

Nathan Lassance, Victor DeMiguel, and Frédéric Vrins. Optimal portfolio diversification via independent component analysis. *Operations Research*, 70(1):55–72, 2022. [7]

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. [2, 7, 9]

Yutong Li, Ruoqing Zhu, Annie Qu, Han Ye, and Zhankun Sun. Topic modeling on triage notes with semiorthogonal nonnegative matrix factorization. *Journal of the American Statistical Association*, pages 1–16, 2021. [7]

H Mann and D Whitney. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Annals of Mathematical Statistics*, 18(1):50–60, 1947. [14]

John Morley. The regulation of mutual fund debt. *Yale Journal on Regulation*, 30: 343, 2013. [6]

Pentti Paatero, Philip K Hopke, Xin-Hua Song, and Ziad Ramadan. Understanding and controlling rotations in factor analytic models. *Chemometrics and Intelligent Laboratory Systems*, 60(1-2):253–264, 2002. [3]

Ioannis Psorakis, Stephen Roberts, Mark Ebden, and Ben Sheldon. Overlapping community detection using bayesian non-negative matrix factorization. *Physical Review E*, 83(6):066114, 2011. [3]

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. [14]

Nornadiah Mohd Razali, Yap Bee Wah, et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011. [15]

Christian P Robert and George Casella. The metropolis—hastings algorithm. In *Monte Carlo statistical methods*, pages 231–283. Springer, 1999. [13]

Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th International Conference on Machine Learning*, pages 880–887, 2008. [3]

Mikkel N Schmidt, Ole Winther, and Lars Kai Hansen. Bayesian non-negative matrix factorization. In *International Conference on Independent Component Analysis and Signal Separation*, pages 540–547. Springer, 2009. [3]

Fritz W Scholz and Michael A Stephens. K-sample anderson–darling tests. *Journal of the American Statistical Association*, 82(399):918–924, 1987. [14]

Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, pages 639–650, 1994. [10]

Chaehee Shin and Maddie White. Fire-sale vulnerabilities of banks: Bank-specific risks under stress and credit drawdowns. *FEDS Notes*, (2020-10):08, 2020. [2]

Hyun Song Shin. Securitisation and financial stability. *The Economic Journal*, 119 (536):309–332, 2009. [5]

Adrian Tobias and Markus K Brunnermeier. Covar. *The American Economic Review*, 106(7):1705, 2016. [2]

David E Upton and Donald S Shannon. The stable paretian distribution, subordinated stochastic processes, and asymptotic lognormality: an empirical investigation. *The Journal of Finance*, 34(4):1031–1039, 1979. [10]

Wolf Wagner. Diversification at financial institutions and systemic crises. *Journal of Financial Intermediation*, 19(3):373–386, 2010. [7]

Dieter Wang, Iman van Lelyveld, and Julia Schaumburg. Do information contagion and business model similarities explain bank credit risk commonalities? *ESRB: Working Paper Series*, 2019. [2, 7]

Yun Yang and David B Dunson. Bayesian conditional tensor factorizations for high-dimensional classification. *Journal of the American Statistical Association*, 111(514):656–669, 2016. [3]

Stavros A Zenios, Andrea Consiglio, Marialena Athanasopoulou, Edmund Moshammer, Angel Gavilan, and Aitor Erce. Risk management for sustainable sovereign debt financing. *Operations Research*, 69(3):755–773, 2021. [32]

Appendix

A. Proofs of Propositions

We denote the rows of a matrix $\boldsymbol{X}_t$ as $\boldsymbol{X}_{i.t}$ and columns as $\boldsymbol{X}_{.jt}$. Also $\boldsymbol{X}_{/x_{i.t}}$ denotes the matrix $\boldsymbol{X}_t$ excluding the i-th row.

Proof of Proposition 1

The following equation expresses the value of bank i assets ($PV_{it} = \sum_k N_{ikt}Y_{kt}$) using W_{ikt} in place of N_{ikt}

$$\sum_k N_{ikt}Y_{kt} = \sum_k W_{ikt}Y_{kt} + R_{it}, \tag{A.1}$$

where the term R_{it} is a remainder term needed for the equality to hold. Isolating the remainder, we have

$$R_{it} = \sum_k N_{ikt}Y_{kt} - \sum_k W_{ikt}Y_{kt}. \tag{A.2}$$

Taking first differences yields

$$\Delta R_{it} = \sum_k (N_{ikt}Y_{kt} - N_{ikt-1}Y_{kt-1}) + (W_{ikt-1}Y_{kt-1} - W_{ikt}Y_{kt}). \tag{A.3}$$

We assume that W_{ikt} is fixed within a short rolling window, which implies that

$$\frac{N_{ikt}Y_{kt}}{\sum_k N_{ikt}Y_{kt}} = \frac{N_{ikt-1}Y_{kt-1}}{\sum_k N_{ikt-1}Y_{kt-1}} \tag{A.4}$$

$$N_{ikt}Y_{kt} = N_{ikt-1}Y_{kt-1}\frac{PV_{it}}{PV_{it-1}}. \tag{A.5}$$

Using the assumption that $W_{ikt} = W_{ikt-1}$, we can write Equation A.3 as

$$\Delta R_{it} = \sum_k (N_{ikt}Y_{kt} - N_{ikt-1}Y_{kt-1}) + (W_{ikt-1}\Delta Y_{kt}) \tag{A.6}$$

$$\Delta R_{it} = \sum_k (N_{ikt-1}Y_{kt-1}\frac{PV_{it}}{PV_{it-1}} - N_{ikt-1}Y_{kt-1}) + (W_{ikt-1}\Delta Y_{kt}) \tag{A.7}$$

$$\Delta R_{it} = \sum_k N_{ikt-1} Y_{kt-1} \left(\frac{PV_{it}}{PV_{it-1}} - 1 \right) + (W_{ikt-1} \Delta Y_{kt}). \tag{A.8}$$

Since bank size tends to not change drastically overnight, the ratio $\frac{PV_{it}}{PV_{it-1}} \approx 1$. When $\Delta Y_{kt} = \Delta W_{ikt} = 0$, then ΔR_{it} will be negligibly small and can be modeled as additive noise. We therefore have that when $\Delta Y_{kt} = \Delta W_{ikt} = 0$,

$$\Delta PV_{it} = \sum_k \Delta W_{ikt} Y_{kt} + \epsilon_{it}, \tag{A.9}$$

where $\epsilon_{it} = \sum_k N_{ikt-1} Y_{kt-1} (\frac{PV_{it}}{PV_{it-1}} - 1)$.

$\square$

Proof of Proposition 2

In the following we drop the subscript t for readability.

Fuzzy K-means aims to minimize the objective function

$$\min_{W_{ik}, \boldsymbol{\mu}_k} \sum_{i=1}^{n} \sum_{k=1}^{K} W_{ik}^2 \| \boldsymbol{Z}_{i.} - \boldsymbol{\mu}_k \|_2^2, \tag{A.10}$$

where $\boldsymbol{W}_{ik}$ are probabilities and $\boldsymbol{\mu}_k$ are cluster centroids. Though the original model, which forced $\boldsymbol{W}$ to be binary, was introduced over a half-century ago, K-means remains widely used in a variety of applications. To see how our model relates to fuzzy K-means, we start by writing $\boldsymbol{Z} = [\boldsymbol{Z}_{1.}, \ldots, \boldsymbol{Z}_{n.}]^T$ and $\boldsymbol{V} = [\boldsymbol{V}_{1.}, \ldots, \boldsymbol{V}_{K.}]^T$. Then we can rewrite the main objective function as

$$\begin{aligned} \| \boldsymbol{Z} - \boldsymbol{W}\boldsymbol{V} \| &= \sum_{i=1}^{n} \| \boldsymbol{Z}_{i.} - \sum_{k=1}^{K} W_{ik} \boldsymbol{V}_{k.} \|_2^2 \\ &= \sum_{i=1}^{n} \| \sum_{k=1}^{K} W_{ik} (\boldsymbol{Z}_{i.} - \boldsymbol{V}_{k.}) \|_2^2 \\ &= \sum_{i=1}^{n} \left[\| \sum_{k=1}^{K} W_{ik}^2 (\boldsymbol{Z}_{i.} - \boldsymbol{V}_{k.}) \|_2^2 + \sum_{k \neq l} W_{ik} W_{il} (\boldsymbol{Z}_{i} - \boldsymbol{V}_{k})^T (\boldsymbol{Z}_{i.} - \boldsymbol{V}_{l.}) \right] \end{aligned} \tag{A.11}$$

Note the first term of Equation A.11 is equivalent to the objective function for fuzzy K-means clustering [Bezdek et al., 1984] with squared probabilities denoting the strength of association between each observation and cluster. In the second term of Equation A.11, if the cluster assignment beliefs are proportional to the distance from the data point to the cluster mean, $W_{ik} \propto \frac{1}{||\boldsymbol{Z}_{i.} - \boldsymbol{V}_{k.}||_2}$, then the second term measures $W_{ik}W_{il}(\boldsymbol{Z}_{i.} - \boldsymbol{V}_{k.})^T(\boldsymbol{Z}_{i.} - \boldsymbol{V}_{l.}) \propto \frac{(\boldsymbol{Z}_{i.} - \boldsymbol{V}_{k.})^T(\boldsymbol{Z}_{i.} - \boldsymbol{V}_{l.})}{\|\boldsymbol{Z}_{i.} - \boldsymbol{V}_{k.}\|_2 \|\boldsymbol{Z}_{i.} - \boldsymbol{V}_{l.}\|_2} =$ cosine(θ). Thus, the proposed model clusters observations (banks) by Euclidean and cosine distance.

□

Proof of Proposition 3

We introduce the following result with proofs available in Theorem 2.3 of Chang and Li [1992], Theorem 1 of Johnston [2010], and Exercise IV.1.3 of Bhatia [1997].

LEMMA A.1. *Let $\boldsymbol{P}$ be an $K \times K$ matrix and $\boldsymbol{x}$ a real-valued vector of length K. Then the following holds: $||\boldsymbol{P}\boldsymbol{x}||_1 = ||\boldsymbol{x}||_1$ if and only if $\boldsymbol{P}$ is signed permutation matrix, i.e., every row and column of $\boldsymbol{P}$ has exactly one non-zero entry, which is either 1 or -1.*

Suppose $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{V} + \boldsymbol{\epsilon_t}$ such that the rows of $\boldsymbol{W}$ satisfy STO (i.e., $||\boldsymbol{W}_{i.}||_1 = 1$ for every row i) and non-negativity. Then if $\boldsymbol{W}\boldsymbol{H}^T$ and $\boldsymbol{H}\boldsymbol{V}$ are another valid solution, $\boldsymbol{W}\boldsymbol{H}^T$ must also satisfy STO. By Lemma A.1, $\boldsymbol{H}$ must then be a signed permutation matrix. Further, due to non-negativity requirement, $\boldsymbol{H}$ must a permutation matrix.

□

Proof of Proposition 4

To show that two variables x and y are conditionally independent, by definition we should show that $p(x, y|z) \propto u_1(x|z)u_2(y|z)$, i.e., we want to show that the posterior distribution (conditioning on data z) can be factorized into a product of two appropriate functions. With our model, this condition with respect to $\boldsymbol{V}_t$ is

$$p(V_{kt}, V_{k't}|\boldsymbol{Z}_t, \boldsymbol{W}_t, \boldsymbol{V}_{/v_{kt}, v_{k't}}, \sigma^2) \propto u_1(V_{kt})u_2(V_{k't}). \tag{A.12}$$

We will show that V_{kt} and $V_{k't}$ are dependent because this condition cannot be satisfied. We start by decomposing the posterior probability

$$p(V_{kt}, V_{k't}|\boldsymbol{Z}_t, \boldsymbol{W}_t, \boldsymbol{V}_{/v_{k't}, v_{kt}}, \sigma^2) \propto p(V_{k't})p(V_{kt})p(\boldsymbol{Z}_t|\boldsymbol{W}_t, \boldsymbol{V}_t, \sigma^2), \tag{A.13}$$

which is obtained through standard application of Bayes rule. Then it's easy to see that the independence condition above is satisfied only when $p(\boldsymbol{Z}_t|\boldsymbol{W}_t, \boldsymbol{V}_t, \sigma^2)$ can itself be factorized into a product of two appropriate functions, like u_1 and u_2 above.

By Equation 6 in the main text,

$$p(\boldsymbol{Z}_t|\boldsymbol{W}_t, \boldsymbol{V}_t, \sigma^2) = \prod_{it} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(Z_{it} - \sum_k W_{ikt}V_{kt})^2}{-2\sigma^2}\right). \tag{A.14}$$

Without loss of generality, assume 2 asset classes so that $\sum_k W_{ikt}V_{kt} = W_{i1}V_{kt} + W_{i2}V_{2t}$. Then note that

$$\exp((Z_{it} - \sum_k W_{ikt}V_{kt})^2) = \exp((Z_{it} - W_{i1t}V_{1t} - W_{i2t}V_{2t})^2) \tag{A.15}$$

$$= \exp(Z_{it}^2 + W_{i1t}^2V_{1t}^2 + W_{i2t}^2V_{2t}^2 - 2Z_{it}W_{i1t}V_{1t} - 2Z_{it}W_{i2t}V_{2t} - 2W_{i1t}V_{1t}W_{i2t}V_{2t}). \tag{A.16}$$

Since it is impossible to write $\exp(2W_{i1t}V_{1t}W_{i2t}V_{2t})$ as a product of two functions with arguments V_{1t} and V_{2t} respectively, the overall posterior likelihood for V_{1t} and V_{2t} also cannot be decomposed as such. Thus, we have established that in general the posterior estimates for V_{kt} and $V_{k't}$ will be correlated, i.e., the estimated returns for different asset classes contained in $\boldsymbol{V}_t$ are dependent.

□

Proof of Proposition 5

Posterior of $\boldsymbol{W}_t$

Since $p(\boldsymbol{W}_t) = \prod_{i=1}^n p(\boldsymbol{W}_{i.t})$ (rows are i.i.d.) and $\boldsymbol{W}_{i.t}$ only affects $\boldsymbol{Z}_{i.t}$, it is easy to see that the posterior of $\boldsymbol{W}_t$ is a product of Gaussian likelihood and a Dirichlet prior

$$p(\boldsymbol{W}_{i.t}|\boldsymbol{Z}_t, \boldsymbol{W}_{t/W_{i.}}, \boldsymbol{V}_t, \sigma^2) \propto p(\boldsymbol{W}_{i.t})p(\boldsymbol{Z}_{i.t}|\boldsymbol{W}_{t/W_{i.}}, \boldsymbol{V}_t, \sigma^2). \tag{A.17}$$

These are not conjugate distributions, which means that we can only compute the posterior distribution's value without characterizing the distribution analytically in closed form. As such, we use the Metropolis Hastings algorithms with a uniform proposal distribution, so that a candidate row $\widetilde{\boldsymbol{W}_{i.t}}$ is generated by moving on the probability simplex randomly around the current state of $\boldsymbol{W}_{i.t}$, i.e., $\widetilde{W_{ikt}} = W_{ikt} + \epsilon$, where ϵ is uniform random noise and $\widetilde{\boldsymbol{W}_{i.t}}$ is subject to probability constraints. Then the candidate row is accepted with probability

$$\min\left(1, \frac{p(\widetilde{\boldsymbol{W}_{i.t}}|\boldsymbol{Z}_t, \boldsymbol{W}_{t/W_{i.}}, \boldsymbol{V}_t, \sigma^2)}{p(\boldsymbol{W}_{i.t}|\boldsymbol{Z}_t, \boldsymbol{W}_{t/W_{i.}}, \boldsymbol{V}_t, \sigma^2)}\right). \tag{A.18}$$

Posterior of $\boldsymbol{V}_t$

We start by decomposing the posterior probability

$$p(V_{kt}|\boldsymbol{Z}_t, \boldsymbol{W}_t, \boldsymbol{V}_{t/v_{kt}}, \sigma^2) \propto p(\boldsymbol{V}_t)p(\boldsymbol{Z}_t|\boldsymbol{W}_t, \boldsymbol{V}_t, \sigma^2) \propto p(V_{kt})p(\boldsymbol{Z}_t|\boldsymbol{W}_t, \boldsymbol{V}_t, \sigma^2). \tag{A.19}$$

Recall that V_{kt} is i.i.d $N(\mu, \sigma_V^2)$. Therefore, the posterior of $\boldsymbol{V}_t$ is a product of a Gaussian prior and Gaussian distribution. Due to these being conjugate distributions, we have the posterior of V_{kt} to be

$$p(V_{kt}|\boldsymbol{Z}_t, \boldsymbol{W}_t, \boldsymbol{V}_{/v_{kt}}, \sigma^2) = \text{Normal}(\mu_p, \sigma_p^2) \tag{A.20}$$

where

$$\sigma_p^2 = \left(\frac{\|\boldsymbol{W}_{.kt}\|_2^2}{\sigma^2} + \frac{1}{\sigma_V^2}\right)^{-1} \tag{A.21}$$

$$\mu_p = \sigma_p^2\left(\frac{\tilde{\mu}\|\boldsymbol{W}_{.kt}\|_2^2}{\sigma^2} + \frac{\mu}{\sigma_V^2}\right) \tag{A.22}$$

$$\tilde{\mu} = \frac{\boldsymbol{Z}_{.kt}^T \boldsymbol{W}_{.kt} - (\boldsymbol{V}_t^T \boldsymbol{W}_t^T)_{t.} \boldsymbol{W}_{.kt} + \|\boldsymbol{W}_{.kt}\|_2^2 V_{kt}}{\|\boldsymbol{W}_{.kt}\|_2^2}. \tag{A.23}$$

Posterior of σ^2

We follow standard arguments to exploit conjugate properties of the Inverse Gamma and Normal distributions.

$$
\begin{aligned}
p(\sigma^2|\boldsymbol{W}_t, \boldsymbol{V}_t, \boldsymbol{Z}_t) &\propto p(\sigma^2)p(\boldsymbol{W}_t, \boldsymbol{V}_t, \boldsymbol{Z}_t|\sigma^2) && \text{(A.24)}\\
&\propto p(\sigma^2)p(\boldsymbol{Z}_t|\boldsymbol{W}_t, \boldsymbol{V}_t, \sigma^2)p(\boldsymbol{W}_t, \boldsymbol{V}_t|\boldsymbol{Z}_t, \sigma^2) && \text{(A.25)}\\
&\propto p(\sigma^2)p(\boldsymbol{Z}_t|\boldsymbol{W}_t, \boldsymbol{V}_t, \sigma^2)P(\boldsymbol{W}_t, \boldsymbol{V}_t) && \text{(A.26)}\\
&\propto p(\sigma^2)p(\boldsymbol{Z}_t|\boldsymbol{W}_t, \boldsymbol{V}_t, \sigma^2) && \text{(A.27)}\\
&\propto \text{Inverse Gamma}(\eta, \theta)\text{Normal}(\boldsymbol{Z}_t|\boldsymbol{W}_t, \boldsymbol{V}_t, \sigma^2). && \text{(A.28)}
\end{aligned}
$$

Then by using results from conjugate distributions, the posterior is

$$
p(\sigma^2|\boldsymbol{W}_t, \boldsymbol{V}_t, \boldsymbol{Z}_t) = \text{Inverse Gamma}(\eta', \theta') \tag{A.29}
$$

where

$$
\begin{aligned}
\eta' &= \eta + \frac{nT}{2} + 1 && \text{(A.30)}\\
\theta' &= \frac{1}{2}\sum_{i,t}(Z_{it} - \sum_{k} W_{ikt}V_{kt})^2 + \theta. && \text{(A.31)}
\end{aligned}
$$

Therefore, we can sample directly in the Gibbs sampler from the posterior conditional distribution Inverse Gamma(η', θ'). □

B. Relationship with SRISK

Supporting the notion that dissimilarity is an early warning indicator due to the coarse asset classes in our data, Figure B.1 shows that a unit increase in similarity is associated with lower future systemic risk levels. This effect is most statistically significant in periods, such as following Dodd-Frank, where the overall banking sector holds generally sound portfolios. The VAR optimal lag specification is based on the Akaike information criterion.

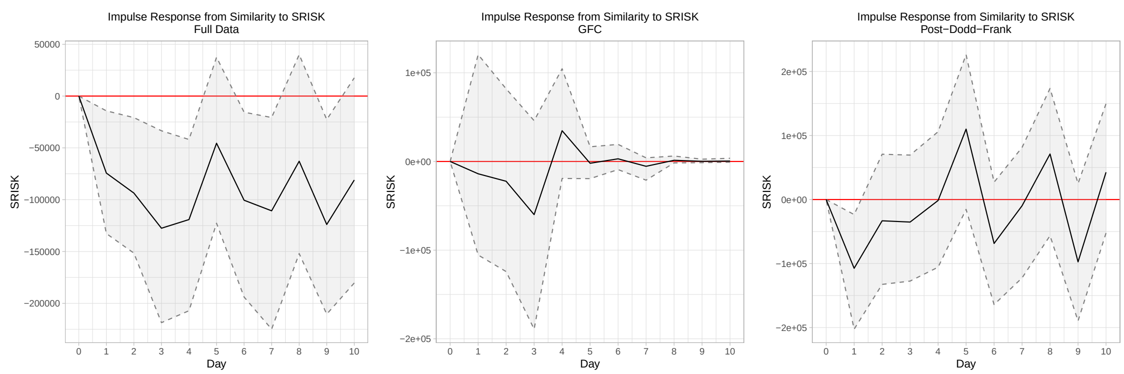

FIGURE B.1. Impulse response functions from the average similarity over all banks to SRISK for the full data (left panel), during the Great Financial Crisis (August 7, 2007 to July 10, 2010; center), and post-Dodd Frank (July 12, 2010 onwards; right). The shaded areas show 95% confidence intervals estimated by bootstrapping.

Rank	Ticker	Name	Status
		Panel A: February 15, 2023	
1	SI	Silvergate Bank	Failed on March 8
2	SIVB	Silicon Valley Bank	Failed on March 10
3	SBNY	Signature Bank	Failed on March 10
4	WAL	Western Alliance Bancorp	Fitch downgraded to BBB- on April 14
5	LOB	Live Oak Bancshares	Shares down over 40% as of May 2023
6	PACW	PacWest Bancorp	Fitch downgraded to BB+ on April 14
7	OZK	Bank OZK	Raised loan loss provisions by 10% in Q1 2023
8	PNFP	Pinnacle Financial Partners Inc.	Moody's downgraded to Baa1 on August 7
9	ZION	Zions Bancorporation	Moody's downgraded to BAA1 on April 21
10	CUBI	Customers Bancorp	Serving crypto customers from SI
		Panel B: March 11, 2023	
Rank	Ticker	Name	Status
1	WAL	Western Alliance Bancorp	Fitch downgraded to BBB- on April 14
2	LOB	Live Oak Bancshares	Shares down over 40% as of May 2023
3	PACW	PacWest Bancorp	Fitch downgraded to BB+ on April 14
4	OZK	Bank OZK	Raised loan loss provisions by 10% in Q1 2023
5	TRMK	Trustmark Corp.	Fitch downgraded to BBB on May 8
6	ZION	Zions Bancorporation	Moody's downgraded to BAA1 on April 21
7	EWBC	East West Bancorp, Inc.	Shares down over 30% as of May 2023
8	COLB	Columbia Banking System Inc	Shares down over 40% as of May 2023
9	CUBI	Customers Bancorp	Serving crypto customers from SI
10	WBS	Webster Financial	Moody's downgraded to Baa1 on August 7
		Panel C: April 28, 2023	
Rank	Ticker	Name	Status
1	LOB	Live Oak Bancshares	Shares down over 40% as of May 2023
2	FRC	First Republic	Acquired by JP Morgan on May 1
3	WAL	Western Alliance Bancorp	Fitch downgraded to BBB- on April 14
4	PACW	PacWest Bancorp	Fitch downgraded to BB+ on April 14
5	CUBI	Customers Bancorp	Serving crypto customers from SI
6	COLB	Columbia Banking System Inc	Shares down over 40% as of May 2023
7	CMA	Comerica Incorporated	Moody's downgraded to Baa1 on April 21
8	BKU	BankUnited	Moody's downgraded to Baa2 on Dec 15
9	UMBF	UMB Financial Corp.	Fitch outlook to negative on May 8
10	TFC	Truist Financial Corporation	Shares down over 30% as of May 2023

TABLE 9. U.S. banks with at least 10 billion dollars in total assets exhibiting tail behavior in their estimated concentration, similarity, and cash holdings.