Economic Quarterly—Volume 101, Number 4—Fourth Quarter 2015—Pages 275–301

Monitoring Economic Activity in Real Time Using Diffusion Indices: Evidence from the Fifth District

Santiago Pinto, Sonya Ravindranath Waddell, Pierre-Daniel G. Sarte

Information on the state of U.S. economic activity compiled by statistical agencies, such as the Bureau of Labor Statistics (BLS), the Bureau of Economic Analysis, or even sections of the Federal Reserve Board, is often released with a one-month lag and is subject to further revisions, typically at the three-month and one-year mark.[1] Moreover, information on economic activity collected by these agencies at a more regional level is limited, so that data on wages, inventories, or shipments at the level of a U.S. state, for example, are not easily obtainable in real time. In part to compensate for this lack of information, several regional Federal Reserve Banks, including Atlanta, Dallas, Kansas City, New York, Philadelphia, and Richmond, produce survey-based diffusion indices that attempt to monitor in real time the direction of change in various regional economic conditions. In this article, we provide an assessment of this effort by the Federal Reserve Bank of Richmond (FRBR) based on ex-post sectoral information related to employment and wages for the Fifth Federal Reserve District. We also provide an assessment of the extent to which more disaggregated regional diffusion indices, not currently constructed but under

We wish to thank R. Andrew Bauer, Jackson Evert, John A. Weinberg, and Zhu Wang for their comments and suggestions. The views expressed in this article are those of the authors and do not necessarily represent those of the Federal Reserve Bank of Richmond or the Federal Reserve System. All errors are our own.

DOI: http://doi.org/10.21144/eq1010401

[1] The work by Croushore and Stark (2001) examines the implications of data revisions on the estimation of macroeconomic models. See Croushore (2011) for a review of the macroeconomic literature on real-time data.

consideration, are likely to inform individual states within the Fifth District.

Diffusion indices of the kind constructed by Federal Reserve Banks as well as many other institutions, such as the widely publicized Institute for Supply Management index (ISM) or the Michigan Survey of Consumers index of consumer sentiment (MSC), aim to measure the breadth of change in a variable of interest, say employment, based on the proportions of its disaggregated component series that move in different directions (increase, decrease, or remain unchanged). This traditional interpretation relying on notions of optimism and pessimism is discussed by Moore (1983) and is distinct from diffusion indices constructed using factor analytic methods in Stock and Watson (2002).

In this article, we build on work by Pinto, Sarte, and Sharp (2015) and highlight the fact that appropriately scaled diffusion indices (defined as the difference between the fractions of sectors that expanded and contracted) capture the contribution of changes in the extensive margin, or the breadth of change, to aggregate changes in a series of interest. For the case of employment in the Fifth Federal Reserve District, we show that changes in this extensive margin, measured by a synthetic diffusion index constructed from observed data, accounts for the bulk of changes in aggregate employment growth. In this context, a synthetic diffusion index is defined as the diffusion index that would be obtained by way of a survey if the sampling were extensive enough to capture the true performance of all sectors making up aggregate employment. Thus, a synthetic diffusion index is a diffusion index that is constructed using disaggregated data that are actually observed ex post.

The finding that a synthetic employment diffusion index for the Fifth District closely follows aggregate employment growth in the District arises in part because aggregate employment growth is well approximated by a formula that uses uniform weights in place of sectoral employment shares in the calculation of the aggregate series. These uniform weights can then naturally be related to the proportion of individual series that move in different directions in a diffusion index. We then show that the actual Fifth District employment diffusion index, produced using firm-level surveys carried out by the FRBR, closely tracks the corresponding ex-post diffusion index constructed using observed data. A key difference is that the survey-based Fifth District index, which proxies closely for aggregate employment growth in the Fifth District (when scaled appropriately), is published in close to real time, whereas the synthetic diffusion index may only be constructed using ex-post data that are subject to revisions up to a year after their initial release.

This article also points to some limitations of using survey-based diffusion indices to track economic changes in real time. In particular, even if a survey-based index were to exactly mimic its "true" synthetic counterpart constructed with data observed ex post, it may perform poorly in tracking the aggregate series of interest. We illustrate this point using a synthetic diffusion index constructed using sectoral data on wages in the Fifth Federal Reserve District. Specifically, we show that such an index fails to effectively track aggregate wage growth in the District. This result follows from the fact that, in the case of wages, changes over time are driven to a greater extent by the intensive margin—the percent change in wages in sectors whose wages are changing in a given month—rather than the extensive margin—the number of sectors whose wages are either increasing or decreasing in a given month. In that sense, the degree to which changes in the extensive margin contribute to changes in an aggregate series is a central consideration in the interpretation of diffusion indices.

Finally, there is a persistent need for timely economic information on U.S. states. Data at the state level are more sparse and less timely than at the national level. At the same time, more granular measures are generally more useful to local economic development practitioners, who tend to be concerned with local information, than measures for the entire Fifth District. Consequently, this article explores some of the implications of producing more localized diffusion indices specific to particular states. To gauge the potential information content of such indices, we examine the behavior of synthetic employment diffusion indices for each of the states within the Fifth Federal Reserve District (District of Columbia, Maryland, North Carolina, South Carolina, Virginia, and West Virginia) constructed using observed data. Our findings suggest that their behavior is far from uniform across indices. Both the volatility of the growth rate in employment and the relative importance of the intensive and extensive margins differ considerably across states. Moreover, the analysis shows that the informational content of the aggregate Fifth District diffusion index would be relevant for states such as Virginia and North Carolina, but much less so for DC and West Virginia. In part, the latter result emerges because economic activity in smaller states such as West Virginia tends to be more concentrated in particular industries or sectors. Thus, in areas where the extensive margin fails to explain a large portion of the overall variation in economic activity, diffusion indices that capture economic information in a larger region do not necessarily provide information that compensates for the lack of real-time economic data on those areas.

This article is organized as follows. Section 1 describes the data used in our analysis. Section 2 reviews key aspects of how aggregate

economic performance relates to economic performance at a more granular level. Section 3 then decomposes economic performance at a disaggregated level into extensive and intensive margins and uses these margins to explain the relationship between diffusion indices and aggregate growth rates. This section also highlights an example in which these measures are closely related, thus providing an underpinning for survey-based diffusion indices designed to capture changes in economic activity in real time. Section 4 highlights some limitations of diffusion indices. Section 5 explores the potential usefulness and other aspects of producing diffusion indices at a more localized level, such as an individual state, rather than an entire Federal Reserve District. Section 6 provides some concluding remarks.

1. DATA

Because diffusion indices aim to provide a sense of the direction of change, or breadth of change, in economic activity, these are most often constructed from disaggregated data such as individual survey data. Diffusion indices constructed using factor analytic methods, as in Stock and Watson (2002), in fact also share this reliance on more granular data. To assess the effectiveness of diffusion indices as real-time estimates of changes in economic activity, this article makes use of two sets of disaggregated data related to the Fifth Federal Reserve District. It also makes use of diffusion indices constructed from surveys of firms in the Fifth Federal Reserve District by the FRBR.

The first set of data is state employment by industry from the Quarterly Census of Employment and Wages (QCEW) program at the BLS. The QCEW data are derived from the quarterly tax reports submitted to state workforce agencies by employers subject to state unemployment insurance laws. Employment covered by the unemployment insurance (UI) programs represents about 97 percent of all wage and salary civilian employment in the country. The employment data are monthly, but are subject to a six-month lag in availability. For most industries, the QCEW data is available from January 1990; therefore the sample used in this analysis covers January 1990 through December 2014. The QCEW data are available at the state level for industries as granular as the six-digit North American Industry Classification (NAICS) code. We include data on the six jurisdictions covered by the Fifth Federal Reserve District (District of Columbia, Maryland, North Carolina, South Carolina, Virginia, and West Virginia) starting at the four-digit NAICS level, subject to data availability for the full time period. To the extent that the data are not available for any industry at the four-digit level, the industry is aggregated to the three-digit NAICS

code, along with all of the other industries covered in that three-digit code. This process is repeated, when necessary, by aggregating the three-digit NAICS into two-digit NAICS codes. Therefore, the final data set is a balanced panel that combines employment by state by industry at the four-, three-, and two-digit NAICS classification levels. Included are data on six regions and 868 industry/state series that are broken down as follows: Washington, D.C. (30 industries), Maryland (157 industries), North Carolina (207 industries), South Carolina (163 industries), Virginia (190 industries), and West Virginia (121 industries).

When the number of establishments in a particular industry in a county or state are too few, the BLS suppresses data in order to preserve confidentiality. Therefore, for certain four- and even three-digit NAICS classification levels, some state data are not available. When the data was combined to create three- or two-digit industries, there were monthly jumps in employment in some of the aggregated industries that represented not an increase or a decrease in employment, but the suppression (or addition) of an industry. For this reason, there were a number of outliers (positive and negative) that created considerable volatility in growth rates. We removed outliers by linearly interpolating growth rates that were above the 90th or below the 10th percentile of the distribution. The data were then seasonally adjusted in SAS using the Census Bureau's X-12 ARIMA program. The adjustment was consistent with the seasonal adjustment that the BLS uses for its Current Employment Statistics payroll employment data.

The second set of data—data on wages—also comes from the QCEW database. The sample period is the same (1990 through 2014); however, the data are available only quarterly. This article uses total wages collected in a state/industry combination over the quarter. The number of industry sectors matches that in the employment data. All manipulation of the data, such as aggregating industries, removing outliers, and controlling for seasonality, is consistent with the employment data.

Finally, the third set of data is collected through the FRBR monthly surveys of manufacturing and service sector activity across the Fifth Federal Reserve District. The survey of manufacturing firms began in 1986 but took its current monthly form in November 1993. The survey asks respondents questions about shipments of finished products, new order volumes, order backlog volumes, capacity utilization (usage of equipment), lead times of suppliers, number of employees, average work week, wages, inventories of finished goods, and expectations of capital expenditures. The services survey began in 1993 and reports on revenues, number of employees, average wages, and prices received. For retailers, the survey also includes questions on current inventory

activity, big ticket sales, and shopper traffic. For this analysis, the manufacturing and services surveys are combined, and diffusion indices are developed from the questions on employment and wages. The survey data are seasonally adjusted according to the same process used to adjust the QCEW data.

There is considerable variation in the number of respondents over time in the Richmond surveys. In 1993, the number of respondents started at around 250 but then fell to a low of 82 respondents by the end of 2000. The number then rose to around 150 respondents by the middle of 2001 and stayed between 150 and 200 respondents until a large jump in 2011 that can be attributed to a consolidation of survey contacts (until 2011, separate surveys were run for the North and South Carolina and Maryland/Washington, D.C., regions). For the past few years, the number of respondents has vacillated around 200 businesses. For wages, the number of respondents jumped considerably from April to May of 1997, since May 1997 was the first month that the question on wages was asked in the manufacturing survey. It is also worth noting that over the years, some questions on the surveys were added, changed, or clarified. Finally, in March 2002, survey respondents began to be able to respond online, although many responses were still faxed and mailed. By December 2010, all responses had to be submitted online.

2. ECONOMIC ACTIVITY IN THE SMALL AND THE LARGE

Formally, diffusion indices are summary statistics of the form,

$$\mu D_t + \kappa, \tag{1}$$

where D_t is the difference between the proportion of a set of disaggregated series that increased between two dates, $t-1$ and t, and the proportion that decreased over the same period,

$$D_t = \frac{N_t^u}{N} - \frac{N_t^d}{N}, \tag{2}$$

where N is the total number of series or categories being considered, say sectors, and N_t^u and N_t^d are the number of series that increased and decreased, respectively; μ and κ are normalizing constants. In the case of the FRBR diffusion indices, $\mu = 100$ and $\kappa = 0$. Thus, index values greater than zero are interpreted as an expansion, say in employment, and negative index values are conversely interpreted as a contraction; upper and lower bounds of 100 and -100 are indicative of all sectors expanding and contracting, respectively. Observe that N_t^u/N in equation (2) also has the interpretation of an average over all categories or sectors where each sector is assigned a value of 1 if

reporting an increase in activity and zero otherwise, and similarly for N_t^d/N.

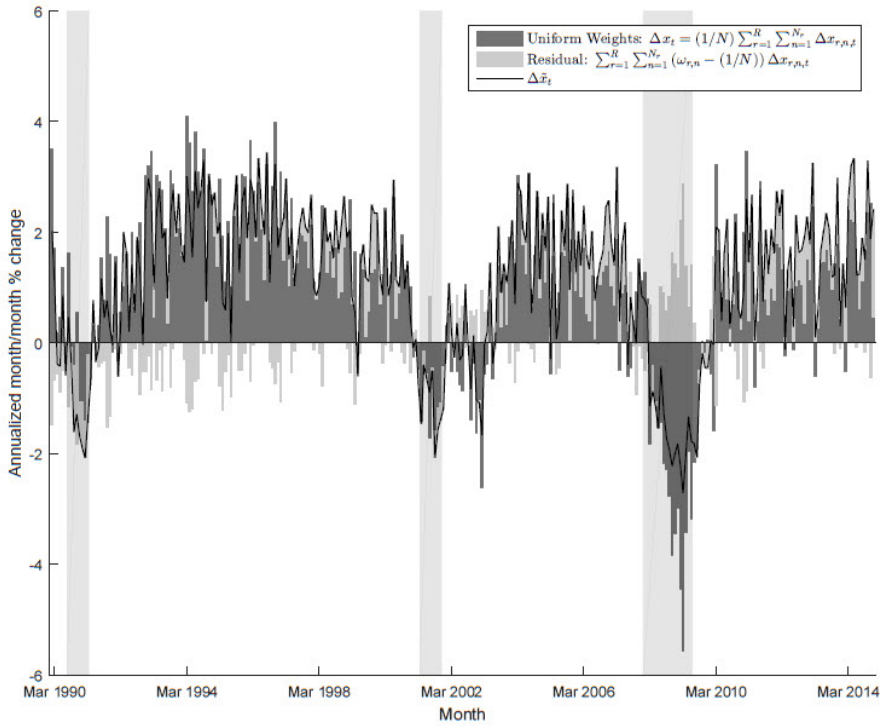
One of the simplest ways in which performance in a given area of the economy is assessed concerns the behavior of the aggregate growth rate in the corresponding variable. Thus, the behavior of aggregate employment growth over a given period, for example, gives us a sense of the performance of the labor market over that period. Aggregate employment growth in turn is a summary of employment growth at a more granular level, say employment growth in the various labor markets across all sectors that make up the aggregate series. In that sense, the estimate of overall growth in a given month hides the details of how this estimate comes about. Put another way, aggregate employment growth may come in moderately high because of a few sectors whose employment grew very rapidly while all other sectors muddled through or even declined or because employment in a wide array of sectors grew at a moderate rate. In contrast, diffusion indices give us a sense of the breadth of economic performance through a summary measure that combines the proportions of sectors whose employment increased relative to those whose employment fell. In this section and the next, we highlight important features of how these different measures of performance relate to each other.

Consider an economy composed of R regions, indexed by $r = 1, ..., R$, in which various sectors operate. There are N_r active sectors in region r, indexed by $n = 1, ..., N_r$. The total number of sectors across all regions is given by $N = \sum_{r=1}^{R} N_r$.[2] Denote employment in a given region r in sector n at date t by $x_{r,n,t}$, and its monthly annualized growth rate by $\Delta x_{n,r,t} = 1200 \times \ln(x_{r,n,t}/x_{r,n,t-1})$. We assume observations over T periods. Because our concern centers around assessing economic conditions in real time, our focus in this paper will be on the highest-frequency data available for a given series, thus monthly in the case of employment. Let $\Delta\widetilde{x}_t$ denote aggregate employment growth across all sectors and regions. Then, it follows by way of an identity that

$$\Delta\widetilde{x}_t = \sum_{r=1}^{R} \sum_{n=1}^{N_r} \omega_{r,n,t} \Delta x_{r,n,t}, \tag{3}$$

[2] For the purpose of our current analysis, the parameter values are given by $R = 6$, $N_{DC} = 30$, $N_{MD} = 157$, $N_{NC} = 207$, $N_{SC} = 163$, $N_{VA} = 190$, $N_{WV} = 121$, and $N = 868$. In the present context, it does not really matter whether sectors are region specific or not. Our analysis relies on aggregate data either at the Fifth District or state level. As we will see later, the aggregation is performed weighting each observation uniformly. Essentially, each sector-region observation is treated as an individual observation.

Figure 1 Employment Growth Rate: Uniform Weights and Residual



where $\omega_{r,n,t} = x_{r,n,t}/x_t$ are weights that, in this case, represent the employment share of a given sector in a given region at time period t. Because the time variation in employment shares is typically small, in the remainder of the paper we consider as a benchmark mean employment shares, $\omega_{r,n}$, independent of time.[3]

Since diffusion indices in (2) implicitly weight individual series uniformly, it is instructive to explore the behavior of a simple aggregate growth rate similar to that in (3) but constructed using uniform weights.[4] In particular, we can write the actual aggregate growth rate,

[3] Foerster, Sarte, and Watson (2011) follow a similar approach.

[4] Observe that N_t^u/N is the sum of series that increase between $t-1$ and t weighted by $1/N$ and similarly for N_t^d/N.

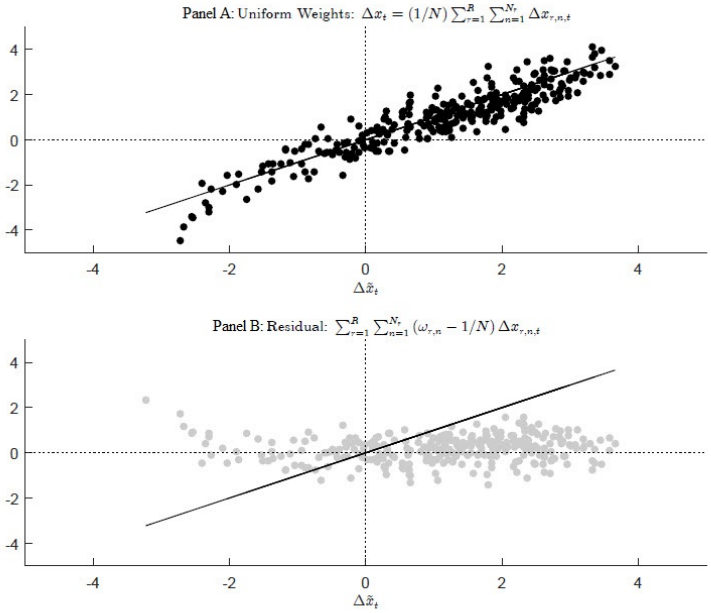
$\Delta\widetilde{x}_t$, as

$$\Delta\widetilde{x}_t = \underbrace{\frac{1}{N}\sum_{r=1}^{R}\sum_{n=1}^{N_r}\Delta x_{r,n,t}}_{\Delta x_t} + \sum_{r=1}^{R}\sum_{n=1}^{N_r}\left(\omega_{r,n} - \frac{1}{N}\right)\Delta x_{r,n,t}, \qquad (4)$$

where $\Delta x_t = \frac{1}{N}\sum_{r=1}^{R}\sum_{n=1}^{N_r}\Delta x_{r,n,t}$ is an approximate growth rate computed using uniform weights, $1/N$, and $\sum_{r=1}^{R}\sum_{n=1}^{N_r}(\omega_{r,n} - \frac{1}{N})\Delta x_{r,n,t}$ is a residual that indicates the importance of deviating from actual shares in using uniform weights to arrive at an aggregate growth rate. Gabaix (2011) refers to the second term on the right-hand side of equation (4) as the granular residual. To provide intuition, in an extreme case where overall performance, $\Delta\widetilde{x}_t$, is mainly determined by a few large sectors in different regions, this second term rather than the first term in equation (4) would tend to dominate the decomposition in (4).

Figure 1 shows the decomposition in equation (4) for employment growth over time across the Fifth Federal Reserve District. By and large, the simple growth rate, Δx_t, represents a good approximation of the actual growth rate, $\Delta\widetilde{x}_t$, throughout the sample period. A notable exception concerns the period covering the Great Recession, when employment fell dramatically and where the simple growth rate and the granular residual moved in opposite directions. However, even in this case, it is the granular residual that moves in a direction opposite the actual growth rate and remains positive throughout the recession while actual aggregate growth is negative. On the whole, Figure 1 suggests that the uniform weighting of the simple growth rate, Δx_t, similar to that of the diffusion indices in (2) whereby each series receives uniform weight $1/N$ conditional on an increase or decrease, has relatively minor implications for measuring overall performance. Figure 2 provides an alternative illustration of the decomposition depicted in equation (4). Specifically, the scatter plot in Figure 2, Panel A, shows that calculations of Δx_t using uniform weights for aggregate employment growth line up closely with actual observations on $\Delta\widetilde{x}_t$ along the 45 degree line. In contrast, the scatter plot in Figure 2, Panel B, depicting the granular residual in (4) is relatively flat with respect to $\Delta\widetilde{x}_t$ around zero. In the Fifth Federal Reserve District, overall employment growth, $\Delta\widetilde{x}_t$, averages to 1.11 percent over our sample period, with the simple aggregate growth rate, Δx_t, averaging 0.94 percent over the same period and the granular residual 0.17 percent. From this point onward, therefore,

Figure 2 Employment Growth Rate: Uniform Weights and Residual



we rely on the simple growth rate, Δx_t, as our benchmark measure of aggregate performance or activity.

3. INTENSIVE AND EXTENSIVE MARGINS OF ECONOMIC ACTIVITY

In order to describe how the diffusion index in (2) for employment, say, and correspondingly aggregate employment growth in (3) are related as summaries of economic activity, let

$$\begin{aligned} \Delta x^{u}_{r,n,t} &= \begin{cases} \Delta x_{r,n,t} & \text{if } \Delta x_{r,n,t} \geq 0 \\ 0 & \text{otherwise} \end{cases} \\ \text{and } \Delta x^{d}_{r,n,t} &= \begin{cases} -\Delta x_{r,n,t} & \text{if } \Delta x_{r,n,t} < 0 \\ 0 & \text{otherwise} \end{cases} . \end{aligned} \tag{5}$$

Simply put, equation (5) distinguishes between those sectors in particular regions that contribute positively to aggregate employment growth, $\Delta x^{u}_{r,n,t}$ (up sectors), and those that reduce aggregate growth,

$\Delta x^d_{r,n,t}$ (down sectors). Then, following Pinto, Sarte, and Sharp (2015), and denoting we may write overall employment growth, Δx_t, in the following way,

$$\Delta x_t = \frac{N^u_t}{N}\mu^u_t - \frac{N^d_t}{N}\mu^d_t, \tag{6}$$

where

$$\mu^a_t = \sum_{r=1}^{R}\sum_{n=1}^{N_r} \Delta x^a_{r,n,t}, \quad a = u, d. \tag{7}$$

In other words, overall growth across all sectors and regions, Δx_t, may be thought of as a weighted sum of average cross-sectional growth rates, where μ^u_t and μ^d_t are the average growth rates of all sectors that add to and subtract from overall growth in a given period, respectively. The weights in (6) are the relative proportions of those sector types.

We can further express each component, $\frac{N^a_t}{N}\mu^a_t$, $a = u, d$, of Δx_t in equation (6) as

$$\begin{aligned}\frac{N^a_t}{N}\mu^a_t &= \mu^a\left(\frac{N^a_t}{N} - \varphi^a\right) + \varphi^a(\mu^a_t - \mu^a) \\ &+ \left(\frac{N^a_t}{N} - \varphi^a\right)(\mu^a_t - \mu^a) + \mu^a\varphi^a, \quad a = u, d,\end{aligned} \tag{8}$$

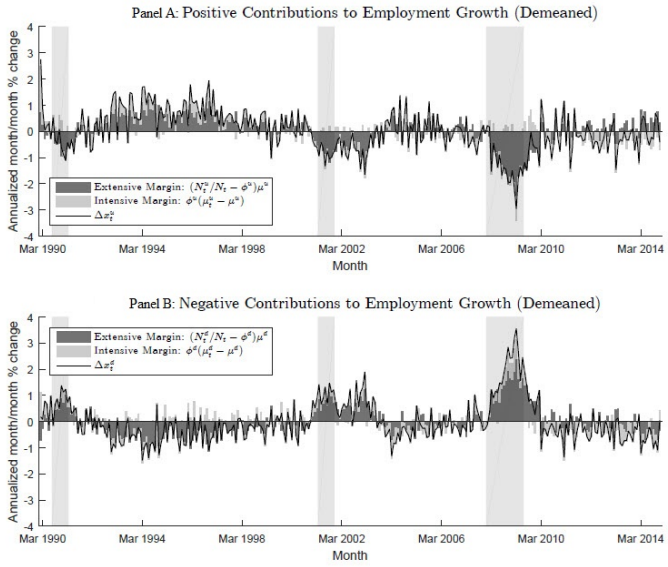
where $\mu^a = \frac{1}{T}\sum_{t=1}^{T}\mu^a_t$, $a = u, d$, are time averages, or long-run cross-sectional averages, of those sectors that contribute positively and negatively to overall growth, and $\varphi^a = \frac{1}{T}\sum_{t=1}^{T}\frac{N^a_t}{N}$, $a = u, d$ are the long-run proportions of those sectors. Thus, equation (8) tells us that, at a point in time, a large increase in overall employment growth by way of $\frac{N^u_t}{N}\mu^u_t$ may come about from the proportion of expanding sectors being higher than usual given their contribution, $\mu^u(\frac{N^u_t}{N} - \varphi^u) > 0$ corresponding to an increasing extensive margin; the cross-sectional average growth rate from those expanding sectors being higher than usual given the typical proportion of those sectors, $\varphi^u(\mu^u_t - \mu^u) > 0$ corresponding to an increasing intensive margin, or both when both are true, $(\frac{N^u_t}{N} - \varphi^u)(\mu^u_t - \mu^u) > 0$. The decline in overall growth by way of $\frac{N^d_t}{N}\mu^d_t$ may be described similarly.[5]

Combining equations (6) and (8), it follows that

$$\Delta x_t \cong \underbrace{\varphi^u(\mu^u_t - \mu^u) - \varphi^d(\mu^d_t - \mu^d)}_{\text{Change in intensive margin}} + \underbrace{\mu^u D_t}_{\text{Change in extensive margin}}, \tag{9}$$

[5] Since the total number of series or sectors is fixed in this context, we should perhaps be referring to a notion of quasi-extensive margin in the sense that entry and exit are not operative.

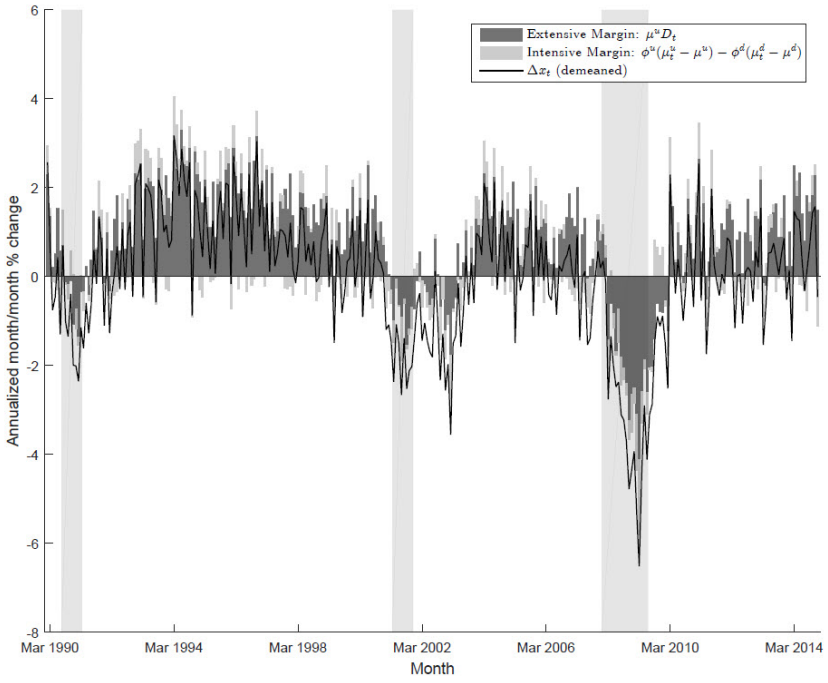
Figure 3 Employment Growth Rate Decomposition: Positive and Negative Contributions



where $D_t = (\frac{N_t^u}{N} - \frac{N_t^d}{N})$ is the difference in the proportions of sectors experiencing positive growth and negative growth respectively defined earlier. In other words, overall economic performance, as measured by an aggregate growth rate, may be interpreted as approximately arising from changes in an intensive margin, the difference between the intensity with which expanding sectors grew and that with which contracting sectors declined, and changes in an extensive margin, the difference between the fractions of sectors that expanded and contracted or the breadth of change in economic activity. The relationship in equation (9) is only approximate in the sense that μ^u and μ^d are close in practice but not necessarily identical so that (9) makes use of the fact that $\mu^d = \mu^u - \delta$ for some small δ. Moreover, the difference in the interaction between intensive and extensive margins in equation (8), $(\frac{N_t^a}{N} - \varphi^a)(\mu_t^a - \mu^a), \quad a = u, d$, may also matter at times.[6] On the whole, however, the question at hand at this point is: which of the two

[6] Refer to Pinto, Sarte, and Sharp (2015) for a complete derivation of expression (9).

Figure 4 Employment Growth Rate Decomposition: Intensive and Extensive Margins



margins in equation (9) tends to explain variations in Δx_t, if any, as a summary of overall economic performance?

Figure 3, Panel A, shows the decomposition of the positive contributions to aggregate employment growth in the Fifth Federal Reserve District, $\frac{N_t^u}{N}\mu_t^u$, in terms of intensive and extensive margins. In this case, $\mu^u = 9.23$, so that expanding sectors contribute about 9.2 percent to employment growth on average, and $\varphi^u = 0.54$, so that expanding sectors represent about 54 percent of all sectors on average. Figure 3, Panel A, makes it clear that variations in $\frac{N_t^u}{N}\mu_t^u$ are to a large degree influenced by variations in the extensive margin. Figure 3, Panel B, shows the decomposition of the negative contributions to aggregate employment growth, $\frac{N_t^d}{N}\mu^d$. As in Figure 3, Panel A, the extensive margin dominates variations in $\frac{N_t^d}{N}\mu^d$. Declining sectors represent about 46 percent of all sectors on average, $\varphi^d = 0.46$, while these sectors reduce aggregate employment growth by 9 percent on average, $\mu^d = 8.95$.

Figure 4 combines Panels A and B of Figure 3 in the manner suggested by equation (9). As expected from the behavior of the individual components $\frac{N_t^u}{N}\mu_t^u$ and $\frac{N_t^d}{N}\mu_t^d$ of Δx_t, changes in the extensive margin explain most of the variations in aggregate employment growth in the Fifth Federal Reserve District. It is interesting to note that in the period following the Great Recession, even though the intensive margin becomes positive, the employment growth rate is still negative. Our analysis reveals that this outcome arises because a large number of sectors are still experiencing a decline in employment (in other words, the extensive margin is still negative), and this effect more than compensates for the positive effect of the intensive margin on the employment growth rate. As a result, the expansion in aggregate employment in the Fifth District since 2009 is largely influenced by the behavior of the extensive margin, or the percentage of sectors experiencing an increase in employment.

The exercises above suggest that, insofar as variations in the extensive margin explain the bulk of aggregate growth in a variable of interest, diffusion indices measuring the breadth of change in economic activity, when appropriately scaled, may serve as a close indication of aggregate growth. Equation (9) makes use of the mean cross-sectional growth rate of expanding sectors, μ^u, to scale the diffusion index. As explained earlier, without much loss of generality, μ^d could also have been used since the two estimates are close. More generally, the scaling factor might be chosen so as to maximize the explanatory power of changes in the extensive margin, D_t, with respect to Δx_t based on ex-post observations. In particular, let D_t^S denote the "true" synthetic diffusion index capturing actual changes in the proportions of expanding and contracting sectors observed ex post. We might think of D_t^S as arising from a survey of firms with a large enough sample to capture the true performance of all sectors making up aggregate employment. Thus, we might then rewrite equation (9) as

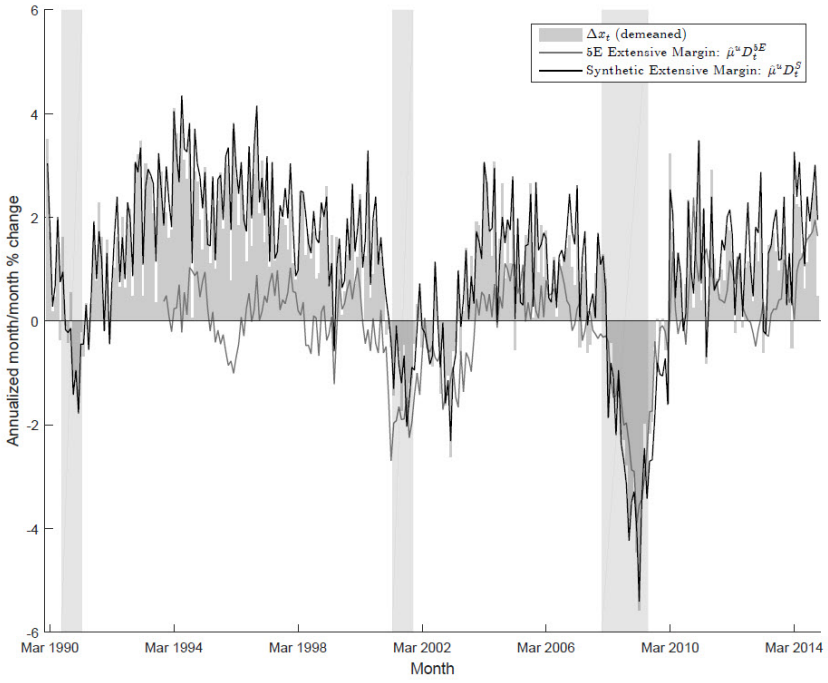
$$\Delta x_t = \underbrace{\alpha + \varepsilon_t}_{\text{Change in intensive margin}} + \underbrace{\mu D_t^S}_{\text{Change in extensive margin}}, \tag{10}$$

and choose α and μ according to a least squares criterion. This yields $\widehat{\mu} = 10.80$ instead of 9.23 used in Figure 4.

Figure 5 illustrates the behavior of the "true" synthetic diffusion index D_t^S, scaled by $\widehat{\mu}$ from (10), against the employment diffusion index produced by the FRBR for the Fifth Federal Reserve District.[7] As discussed earlier, this employment diffusion index represents the

[7] When μ is estimated using the diffusion index calculated by the FRBR instead of D_t^S in equation (10), we obtain $\widehat{\mu} = 12.10$.

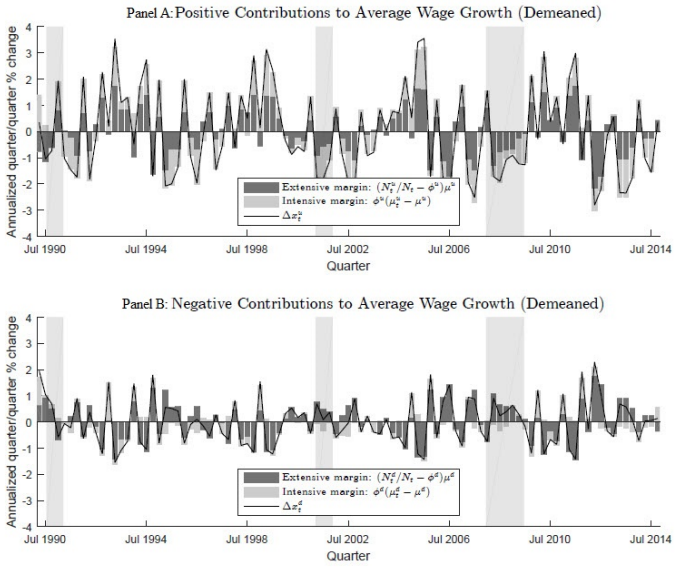
Figure 5 Employment Growth Rate: Synthetic and FRBR Extensive Margins



share of respondents to the FRBR manufacturing and service sector surveys who reported increased employment in the last month minus the share of respondents who reported decreased employment. The data for the month are collected from respondents through the third Wednesday of every month and are available publicly on the fourth Tuesday of every month; thus, they are the timeliest regional data available.

As indicated by Figure 5, the survey-based diffusion index produced in real time by the FRBR lines up remarkably closely with the synthetic diffusion index produced from employment data observed ex post. Figure 5 also shows, however, that the performance of the FRBR employment index improves over time. In the early years of the FRBR's index, from 1993 to 2001, the survey-based diffusion index and the synthetic index are somewhat far apart. There have been a few changes to the surveys over the years. As is clear from the earlier discussion, the number of respondents and the sampling of the respondents changed

Figure 6 Wage Growth Rate Decomposition: Positive and Negative Contributions



over the years due to both economic changes in the region and changes to the survey process. One such process change was that in March 2002, survey respondents began to be able to respond online, although many responses were still faxed and mailed. By December 2010, all responses had to be submitted online. Thus, with those changes, the Richmond Fed's employment diffusion index begins to track its synthetic counterpart much more closely beginning in 2002. Between June 2002 and December 2014, the correlation between the survey-based diffusion index and the synthetic index constructed from observed data is 0.77. In addition to the changes in the survey process explained earlier, other reasons, including variations in the survey composition and sample size, may also explain the shift observed in the survey series starting in 2002.[8]

[8] A more detailed analysis is required to identify such factors. We will revisit this issue in future work.

4. LIMITATIONS OF DIFFUSION INDICES

As the previous section suggests, the importance of the link between diffusion indices and aggregate growth rates hinges crucially on the relative contribution of the extensive margin of activity to overall growth. In the case of changes in employment in the Fifth Federal Reserve District, we saw that changes in the extensive margin contributed significantly to overall employment growth. There is nothing to suggest, however, that this should be the case for all aggregate series of interest. To highlight the potential limitations of diffusion indices, we consider an effort to track wage pressures in real time by way of changes in the extensive margin that keeps track of the proportion of sectors that are seeing increases and decreases in average wages.

One natural definition of an overall average wage that takes into account wages in all sectors and regions is given by

$$\widetilde{w}_t = \sum_{r=1}^{R} \sum_{n=1}^{N_r} \frac{x_{r,n}}{x} w_{r,n,t} \tag{11}$$

where $w_{r,n,t}$ is the average wage in region r in sector n at date t, and $\frac{x_{r,n}}{x}$ is the corresponding mean employment share in that region and sector. Average wage growth, $\Delta\widetilde{w}_t$, then follows approximately

$$\Delta\widetilde{w}_t = \sum_{r=1}^{R} \sum_{n=1}^{N_r} \frac{W_{r,n}}{W} \Delta w_{r,n,t}, \tag{12}$$

where $W_{r,n}$ is the (mean) total wage bill in region r in sector n, and W is the (mean) total wage bill across all sectors and regions. As in the previous section, we can decompose average wage growth in equation (12) into a uniformly weighted growth rate and a granular residual,

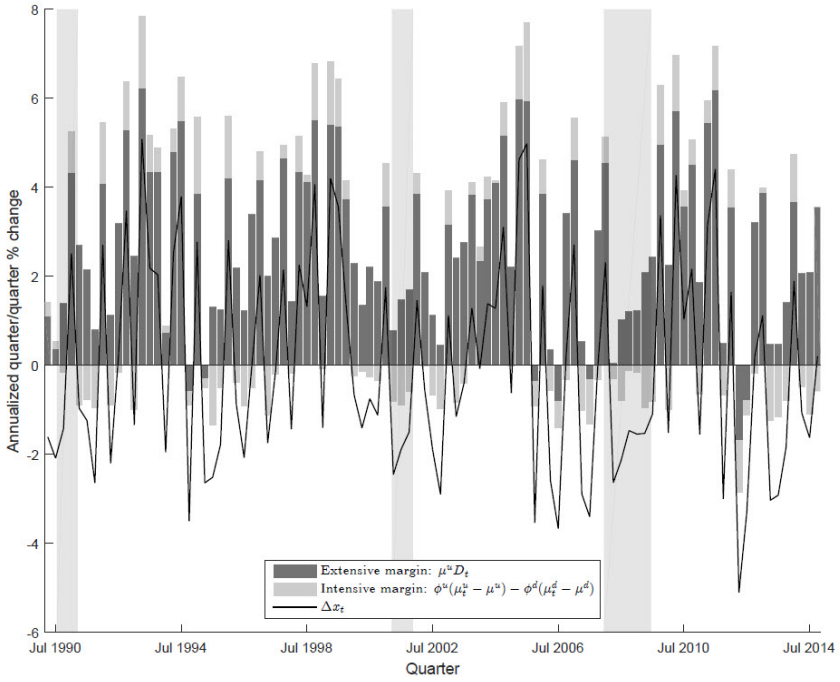
$$\Delta\widetilde{w}_t = \underbrace{\frac{1}{N} \sum_{r=1}^{R} \sum_{n=1}^{N_r} \Delta w_{r,n,t}}_{\Delta w_t} + \sum_{r=1}^{R} \sum_{n=1}^{N_r} \left(\frac{W_{r,n}}{W} - \frac{1}{N} \right) \Delta w_{r,n,t}, \tag{13}$$

and further decompose the simple average growth rate, Δw_t, into intensive and extensive margin changes,

$$\Delta w_t \cong \left[\varphi^u (\mu_t^u - \mu^u) - \varphi^d (\mu_t^d - \mu^d) \right] + \mu^u D_t. \tag{14}$$

Analogously to the decomposition of employment in the previous section, changes in the intensive margin, $[\varphi^u(\mu_t^u - \mu^u) - \varphi^d(\mu_t^d - \mu^d)]$, capture how high increasing wages are rising relative to how badly declining wages are falling in those sectors and regions where wages

Figure 7 Wage Growth Rate Decomposition: Intensive and Extensive Margins

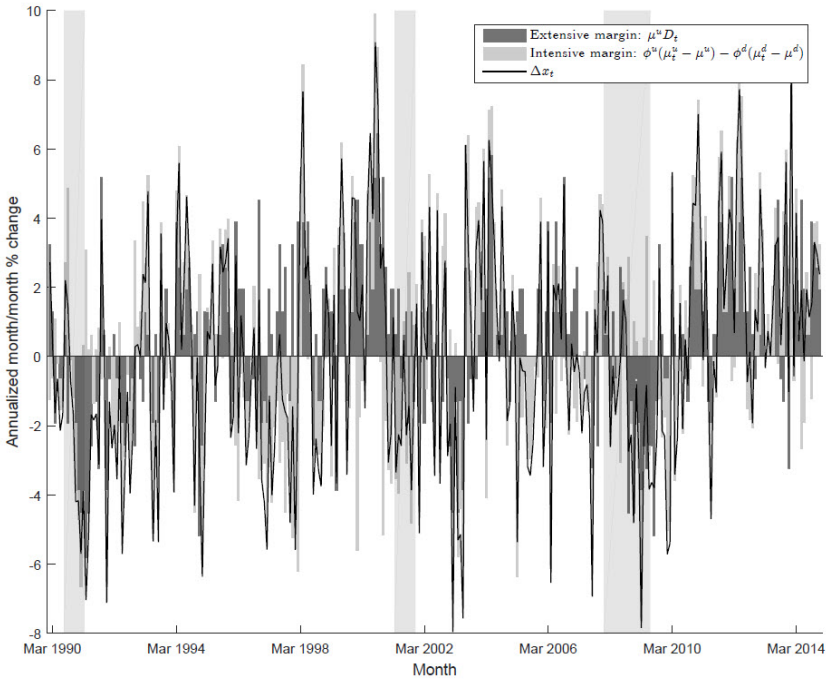


are changing. Changes in the extensive margin, $\mu^u D_t$, reflect the extent to which increasing wages are widespread across sectors relative to decreasing wages.

Figure 6 shows the decomposition of the positive and negative contributions to average wage growth into their respective intensive margins, $\varphi^a(\mu_t^a - \mu^a)$, and extensive margins, $\mu^a \left(\frac{N_t^a}{N} - \varphi^a\right)$, $a = u, d$. A salient feature of the positive contributions is that the intensive margin is at least as important at explaining $\frac{N_t^u}{N}\mu_t^u$ as the extensive margin, with periods in which the former even dominates the latter. The negative contributions to the average wage growth rate are, however, generally much smaller and mostly dominated by the extensive margin.

Figure 7 shows the overall decomposition of the growth rate in average wages into the intensive and extensive margins, as indicated by (14). From the figure, we observe that changes in the extensive margin are, with only a few exceptions, always positive. The (demeaned)

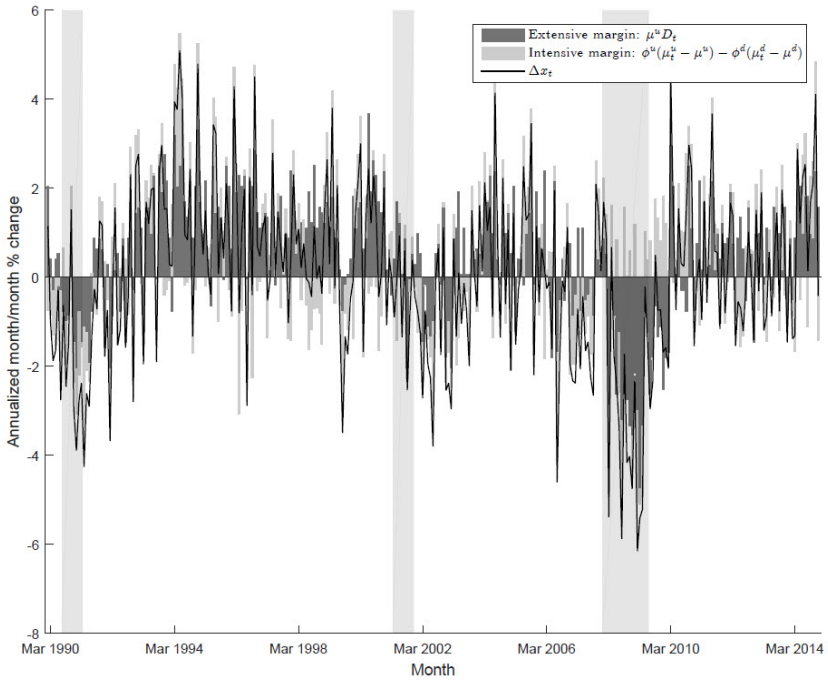
Figure 8 DC - Employment Growth Rate Decomposition: Intensive and Extensive Margins



growth rate of average wages, however, is frequently negative, generally coinciding with periods in which changes in the intensive margin are negative. The decomposition in (14) reveals that the intensive margin will tend to be negative whenever the average wage growth rate of sectors reporting an increase at time period t, μ_t^u, becomes small relative to the corresponding growth rate of those reporting a decline, μ_t^d. In our sample, the changes in μ_t^u dominate, especially in time periods in which the economic activity is low or declining. For instance, average wage growth is below its mean from the first quarter of 2008 until the second quarter of 2009, coinciding with a time period in which μ_t^u was also below its mean. The behavior of μ_t^d is, however, more erratic, with quarters in which μ_t^d was even below its mean during that same period.[9]

[9] Also, note that while the correlation between μ_t^u and Δw_t is 0.80, the correlation between μ_t^d and Δw_t is -0.14.

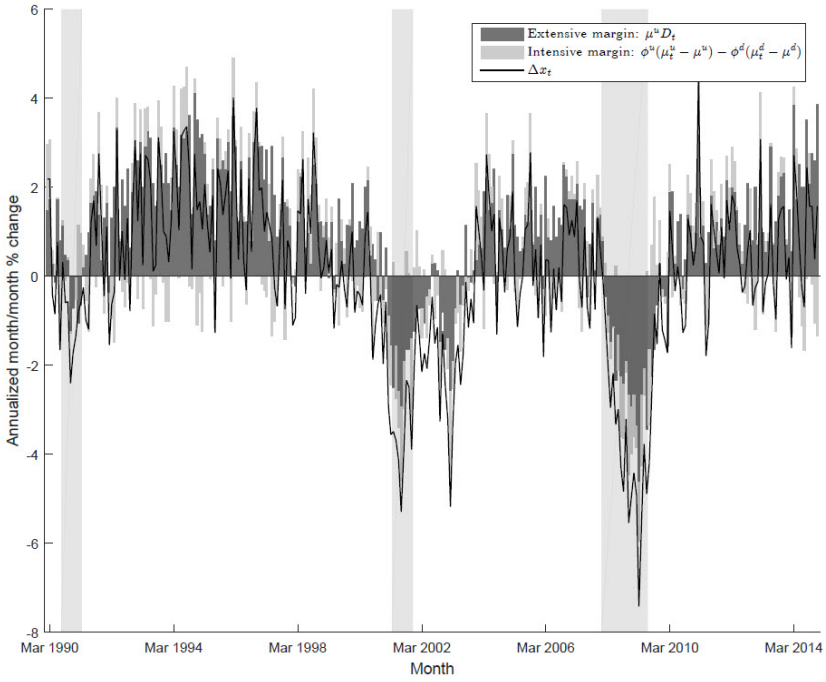
Figure 9 MD - Employment Growth Rate Decomposition: Intensive and Extensive Margins



Put another way, in the case of average wages, changes in the extensive margin are frequently at odds with the behavior of its overall growth rate, highlighting the limitations of diffusion indices as real-time indicators of economic activity. One reason that explains the weak connection between the extensive margin and the overall growth rate in average wages is that wages seldom decline in nominal terms. Other series may certainly show the same kind of pattern.[10]

[10] Further examination of the underlying factors explaining the behavior of different series of interest (specifically those series included in the FRBR survey) would allow us to determine which ones are more like the employment series, where the extensive margin plays a dominant role, and which ones share more closely the characteristics of the wage series.

Figure 10 NC - Employment Growth Rate Decomposition: Intensive and Extensive Margins

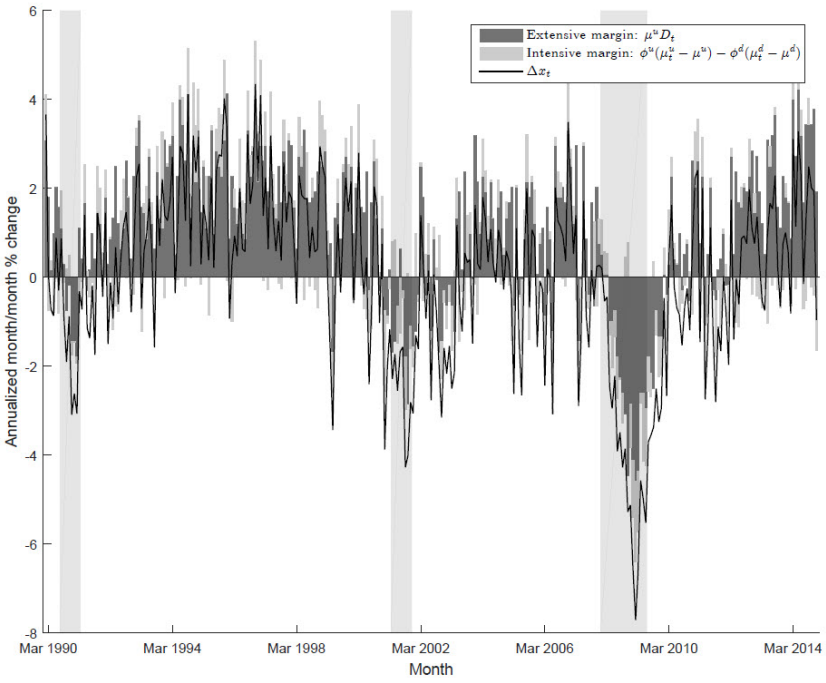


5. STATE-LEVEL DIFFUSION INDICES

Although the survey-based diffusion index for the Fifth District aids in understanding economic activity for the entire region, the dearth of data available for individual states combined with the important role that state boundaries play in economic activity and policymaking, mean that measures of activity at the state level would be more useful to many local policymakers or economic development practitioners than measures related to the Fifth Federal Reserve District. The manufacturing and service sector surveys provide information that is not otherwise available at the state level (such as new orders of manufactured goods or retail shopper traffic) in a timely fashion and include respondents' projections of future activity.

Mainly due to data limitations, the FRBR is unable to construct and report state-level diffusion indices for manufacturing and services

Figure 11 SC - Employment Growth Rate Decomposition: Intensive and Extensive Margins



separately.[11] The FRBR combines survey responses from every state and calculates a Fifth District diffusion index. However, the evolution of this aggregate indicator may not accurately track the performance of each individual state in the District. To understand the implications of conducting an analysis at the level of the District rather than individual states, we use state employment data from QCEW and apply the methodology introduced in Section 3 to each of the states in the Fifth District. Specifically, we decompose state employment growth rates into their intensive and extensive margins and construct synthetic state-level employment diffusion indices.

Figures 8 through 13 show the evolution of the employment growth rate and changes in the intensive and extensive margins for each state in the Fifth District. Table 1 presents their standard deviations, and

[11] However, the FRBR conducts a survey of general business activity for the Carolinas and for Maryland.

Table 1 Employment: Standard Deviation

	$\Delta \mathbf{x}_{r,t}$	Intensive Margin	Extensive Margin
5E	1.52	0.45	1.24
DC	3.26	2.44	2.23
MD	1.98	1.14	1.42
NC	1.92	0.81	1.55
SC	2.03	0.74	1.67
VA	1.79	0.85	1.36
WV	2.02	1.21	1.36

Table 2 the cross-correlations between the calculated synthetic diffusion indices (or extensive margins). A few remarks are worth making. First, the volatility of the employment growth rate differs considerably across states. The standard deviation of $\Delta x_{r,t}$ throughout the period under consideration is almost twice as high in DC (3.26) as it is in Virginia (1.79). Second, the relative importance of intensive and extensive margins in explaining state-level employment growth also differs considerably across states. While changes in the extensive margin explain the bulk of variations in state employment growth in North Carolina, South Carolina and Virginia, they seem much less relevant to employment growth in DC, Maryland, and West Virginia, where the intensive and extensive margins play essentially similar roles. For the states in the Fifth District, economic activity tends to be concentrated in a lower number of sectors in smaller states, with the extensive margin thus becoming relatively less important. Third, the correlation between the synthetic Fifth District diffusion index D_t^S calculated earlier and the state-level diffusion indices also differs across states, as suggested by Table 2.

Table 2 Correlation Matrix: State and Fifth District Diffusion Indices (Extensive Margin)

	5E	**DC**	**MD**	**NC**	**SC**	**VA**	**WV**
5E	1.0000						
DC	0.4450	1.0000					
MD	0.8174	0.4221	1.0000				
NC	0.8969	0.2924	0.6184	1.0000			
SC	0.8511	0.3197	0.5878	0.7580	1.0000		
VA	0.9095	0.3881	0.7364	0.7400	0.7110	1.0000	
WV	0.5443	0.1884	0.3823	0.4122	0.2816	0.4673	1.0000

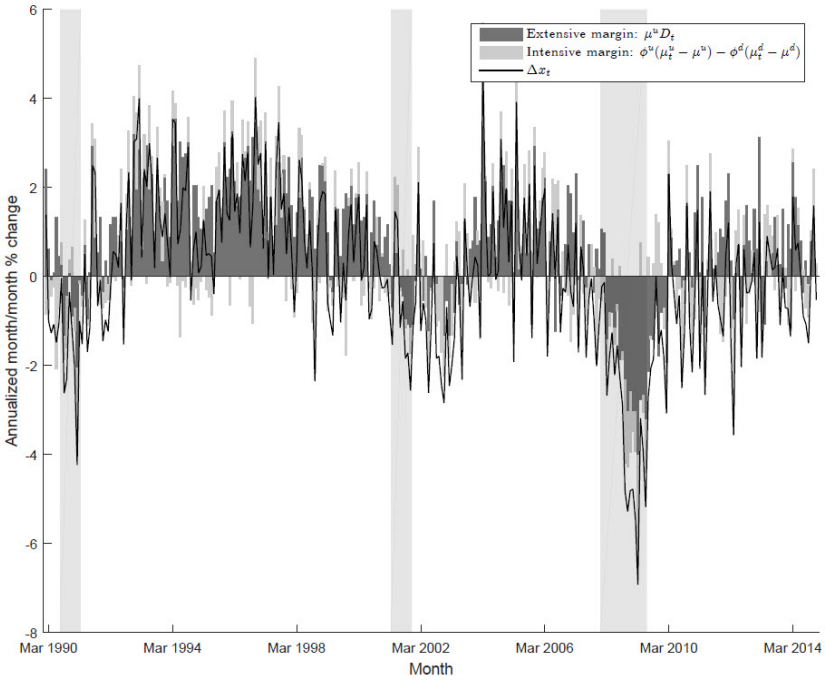
In particular, the diffusion indices for Virginia and North Carolina seem to closely follow the performance of D_t^S, with correlation coefficients of about 0.90. The correlation is also relatively high for South Carolina and Maryland (0.85 and 0.82, respectively). However, the correlations between state and Fifth District indices are much lower for DC and West Virginia (0.45 and 0.54, respectively). Thus, in regions where the extensive margin fails to explain a large component of the overall variation in economic activity, broader-based diffusion indices capturing economic information in surrounding regions do not necessarily make up for the lack of real-time information. Even though the synthetic diffusion index may not accurately represent the behavior of aggregate growth in states where economic activity is concentrated in a few sectors (such as West Virginia), an index based on a large enough sample of survey respondents may perform satisfactorily.

6. CONCLUDING REMARKS

In this article, we provide an analysis of diffusion indices that parses out the conditions under which they are likely to serve as reliable real-time indicators of economic activity. In particular, building on Pinto, Sarte, and Sharp (2015), we highlight the fact that diffusion indices, appropriately scaled, capture the contribution of changes in the extensive margin to aggregate changes in a series of interest. For the case of employment in the Fifth District, we show that changes in this margin in fact account for the bulk of changes in aggregate employment growth.

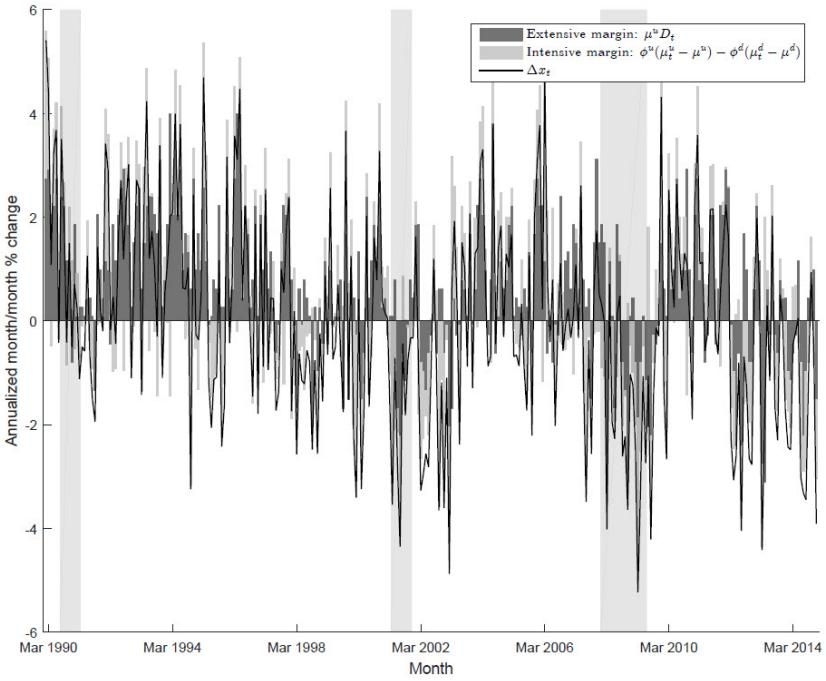
This article also highlights the potential limitations of diffusion indices. Specifically, since diffusion indices capture changes in an extensive margin, these indices are of limited usefulness in cases where aggregate changes are driven by the intensive margin. That is, the intensity with which economic activity increases in particular sectors, for

Figure 12 VA - Employment Growth Rate Decomposition: Intensive and Extensive Margins



example, rather than the number of sectors in which economic activity increases. In the case of average wages, for example, we show that changes in the extensive margin are frequently opposite that its overall growth rate. Finally, we explore the potential usefulness and other aspects of producing diffusion indices at a more localized level, such as an individual state, rather than an entire Federal Reserve District. Given that economic activity is typically more concentrated across sectors in smaller states, and changes in the extensive margin play a smaller role, relying on broader diffusion indices capturing activity in surrounding regions remains of limited use for such states.

Figure 13 WV - Employment Growth Rate Decomposition: Intensive and Extensive Margins



REFERENCES

Croushore, Dean. 2011. "Frontiers of Real-Time Data Analysis." *Journal of Economic Literature* 49 (March): 72–100.

Croushore, Dean, and Tom Stark. 2001. "A Real-Time Data Set for Macroeconomists." *Journal of Econometrics* 105 (November): 111–30.

Foerster, Andrew T., Pierre-Daniel G. Sarte, and Mark W. Watson. 2011. "Sectoral versus Aggregate Shocks: A Structural Factor Analysis of Industrial Production." *Journal of Political Economy* 119 (February): 1–38.

Gabaix, Xavier. 2011. "The Granular Origins of Aggregate Fluctuations." *Econometrica* 79 (May): 733–72.

Moore, Geoffrey H. 1983. "Why the Leading Indicators Really Do Lead." In *Business Cycles, Inflation, and Forecasting*, 2nd ed. Cambridge, Mass.: Ballinger, 339–52.

Pinto, Santiago M., Pierre-Daniel G. Sarte, and Robert Sharp. 2015. "Learning About Consumer Uncertainty from Qualitative Surveys: As Uncertain As Ever." FRB Richmond Working Paper 15-09 (August).

Stock, James H., and Mark W.Watson. 2002. "Macroeconomic Forecasting Using Diffusion Indexes." *Journal of Business and Economic Statistics* 20 (April): 147–62.

Economic Quarterly—Volume 101, Number 4—Fourth Quarter 2015—Pages 303–317

Does Bank Lending Matter for Large Firms' Investment?

Marios Karabarbounis

1. INTRODUCTION

Does bank lending matter for corporate investment? On the one hand, if corporations have easy access to alternative sources of finance such as internal financing, external equity, or bond issuance, then investment will be less affected by how much banks are willing to lend. On the other hand, if corporations are strongly attached to bank lending, then disruptions in bank financing might affect firms' investment.

Starting from Kashyap, Stein, and Wilcox (1993), this question has spurred a large literature.[1] Most studies are subject to the criticism of being unable to distinguish between pure supply variations in bank lending and changes in credit demand. However, the increasing trend of focusing away from macro-level to firm-level data has offered new opportunities to deal with this endogeneity. For example, in a recent article, Chodorow-Reich (2014) used cross-sectional variation in disruptions of banking relationships to analyze the employment effects of the recent financial crisis. His findings point toward significant effects of bank lending for the employment of small firms.

This article uses similar identification techniques to address whether bank lending matters for corporate investment. To my knowledge, there is no work employing microdata on banking relationships to analyze the

For useful comments I thank Bob Hetzel, David Min, Nico Trachter, and John Weinberg. Contact information: marios.karabarbounis@rich.frb.org.

DOI: http://doi.org/10.21144/eq1010402

[1] Other significant papers analyzing the effect of bank lending are Bernanke and Blinder (1988) and Ramey (1993).

effect of bank lending on firm investment. The exercise combines income statement and balance sheet information on publicly listed firms from Compustat with information from Loan Pricing Corporation's DealScan. Following Chodorow-Reich (2014), I use DealScan data to identify the banking institutions in lending relationships with the firms in the Compustat sample. For each bank, I construct an index—the bank lending ratio—summarizing how much banks decreased lending after the crisis compared to their pre-recession level. I then construct a firm-specific measure of bank lending supply: the relative exposure of each firm to banks that faced severe lending disruptions. Intuitively, a firm heavily borrowing from a bank that experienced difficulties would find it harder to expand its credit compared with a firm that was borrowing from healthier banks.

The key idea is that disruptions in credit could be considered an exogenous event for a particular firm. For example, banks that experienced financial turmoil did so mainly due to their exposure to risky financial instruments such as toxic mortgage loans. Using this type of variation, one can abstract from traditional measures of bank lending that are more likely to suffer from endogeneity. An example of such measure is the aggregate bank share of debt issuance (Kashyap, Stein, and Wilcox 1993).

It turns out that the two measures yield completely different results. The aggregate bank share is strongly correlated with the change in investment. During periods of lower bank share, firm-level investment decreases. In sharp contrast, our "exposure" measure (a proxy for a firm's ability to borrow) does not affect investment in a significant way.

A caveat of our exercise is that we focus on publicly listed firms from Compustat. These firms are typically large firms that can substitute more easily bank lending with not only external equity financing but also internal equity. As a result, it would be a mistake to extrapolate our findings for the universe of U.S. firms. It is very likely that bank lending can have significant effects on smaller firms, which are not included in the sample.

This paper contributes to the literature analyzing the effect of bank lending on macroeconomic variables. Bernanke and Blinder (1988) develop a model that allows roles for both money and bank loans. Ramey (1993) studies the importance of the credit channel on the transmission of monetary policy. Kashyap, Stein, and Wilcox (1993) explore the existence of a loan supply channel using bank loan and commercial paper measures.

Berger and Udell (1995) show that small firms with longer banking relationships borrow at lower rates and are less likely to pledge collateral than other small firms. Ivashina and Scharfstein (2010) show that

banks cut their lending less if they were not reliant on short-term debt and had better access to deposit financing. Jiminez, Mian, Peydro, and Saurina (2014) analyze the impact of securitization of real estate assets on the supply of credit to non-real estate firms. Becker and Ivashina (2014) also use firm-level evidence from DealScan. While their main focus is to provide evidence of bank supply shocks, they also related the aggregate bank share to investment. As mentioned, we consider this measure to be prone to endogeneity. Hence, this paper exploits a different measure based on bank lending relationships.

2. EMPIRICAL ANALYSIS

Data Description

To analyze the effect of bank lending on investment, we combine two datasets. The first is the Compustat annual database, which includes balance sheet information on publicly listed companies. Since these companies are much larger than the representative firm, our analysis is better viewed as applying to large firms. The second dataset is the Loan Pricing Corporation's DealScan from Thomson Reuters. This dataset includes daily information on new bank loan issuances for a large set of companies both private and public. The information on loan characteristics includes (among others) the name of the firm undertaking the loan, the amount issued, the issue date, the type and purpose of the loan, and the cost and maturity of the loan. Moreover, there is information on the name of the banks that act as a syndicate to lend money as well as which bank(s) act as book manager (leader of deal). Being able to identify where the loan originates is crucial for the analysis.

We will focus only on nonfinancial U.S. firms for the period between 2000–13. Investment is defined as capital expenditures on property, plant, and equipment (Compustat data item #30). Within DealScan, I exclude firms in financial- and government-affiliated industries and only include loans used for construction of capital buildings or other construction, capital expenditures, and property development. This way I exclude loan deals not used for real investment purposes such as refinancing, stock buyback, or mergers. We deflate all variables by the Producer Price Index.

After these restrictions, we are left with a total of 2,022 firms and a total of 11,390 observations. As mentioned, the DealScan sample includes a much larger set of firms both private and public. In particular, it includes 21,457 firms and a total of 114,989 observations. Table 1 provides summary statistics for loan issuance. We report these statistics for both our sample (the intersection of Compustat and DealScan)

Table 1 Summary Statistics for Loan Issuances (2000–13)

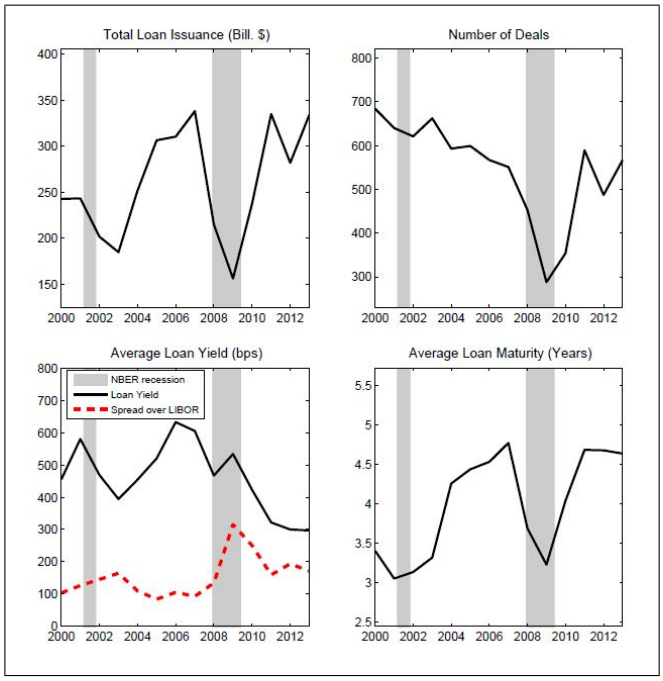
	DealScan	**DealScan & Compustat**
Deals #	29,447	7,670
Average Amount (Millions $)	168.9	263.6
Maturity (Years)	3.8	3.8
Spread over LIBOR (bps)	166.4	132.4
Firms #	21,457	2,022
Observations #	114,989	11,390

and the complete DealScan dataset. During our period, there are a total of 7,670 loans issued to Compustat firms. The total number of loan deals in all firms in DealScan is 29,447. The average amount of a loan deal is $263 million in our sample. In the full DealScan dataset, the average amount is $169 million. In both, the average loan deal matures in 3.8 years. We measure the cost of a loan deal as the spread over the LIBOR of the respective maturity. To compute the average, we weight each deal by its size relative to the total amount issued in the given year. In our sample the average spread is 132 basis points. In DealScan it is higher, around 166 basis points.

Figure 1 plots several patterns of bank loan financing during 2000–13. The most striking pattern is the sharp reduction in bank loan issuance during the recent crisis. Issuance decreased from $215 billion in 2007 to $156 billion in 2009 (upper left panel). By 2011, bank lending had returned to the pre-recession levels. The procyclicality of bank financing is also evident in the 2001 recession. The upper right panel plots the number of loan deals per year. The number decreased from 551 in 2007 to 288 in 2009, almost one-half of the pre-recession level. I also compute the average amount per loan deal, although it is not plotted in Figure 1. The per-deal amount also decreased from $338 million in 2007 to $156 million in 2009. Hence, the sharp decline in loan financing was the result of both fewer firms getting a loan and of those that borrowed less.

In parallel with the decline in loan financing, the cost of loans rose sharply. The lower left panel of Figure 1 plots the average yield as the spread over LIBOR and the loan yield, which is defined as the spread plus LIBOR. The difference between the two lines gives the LIBOR path. As mentioned, deals are weighted by their size. Loan spreads increased from 92 bps in 2007 to 315 bps in 2009. Although the spreads decreased in 2010, they stabilized at a higher level compared with the pre-recession level. However, the overall yield did not increase as much due to the decreasing interest rates of LIBOR. In 2013, the yield was significantly lower than the pre-recession level. Finally, the lower right

Figure 1 Loan Issuance



Notes: Upper left panel shows the total loan issuance in billions of dollars. Upper right panel shows the total number of loan deals. All amounts are deflated using the PPI. Lower left panel shows the yield to maturity in BPS. Dotted line shows the spread over LIBOR, while the solid shows spread + LIBOR. Lower right panel shows the average maturity of loan deals.

panel of Figure 1 plots the average maturity of loan deals in our sample, which decreased from 4.7 years in 2007 to 3.2 years in 2009.

Note that the patterns outlined above seem to hold for the 2001 recession as well. Total loan issuance and number of deals decreased (but not as sharply). The loan yield decreased, but the spread over LIBOR increased. The only difference is that average loan maturity was increasing from a low rate even from 2001 and accelerated once the recession was over.

The Identification Scheme

Our main goal is to understand how variations in bank loan supply affect the firms' investment decisions. A simple approach is to regress the

change in investment by firm i in period t on some aggregate measure of bank loan supply in period t:

$$\Delta\text{Investment}_{i,t} \;=\; \beta_0 \;+\; \beta_1 \text{ Bank Loan Supply}_t + \varepsilon_{i,t}$$

The coefficient β_1 gives the causal effect of the change in firms' investment due to changes in banks' loan supply if there are no underlying factors affecting both variables. Hence, the identification assumption is that $Cov(\text{Bank Loan Supply}_t, \varepsilon_{i,t}) = 0$. This is a strong assumption that may very likely be violated. For example, changes in both investment and bank loan supply may be driven by business cycle conditions. In particular, firms may decrease their investment due to lower expected demand and consequently decrease their demand for credit. Hence, investment may be responsible for the decrease in bank lending, not the other way around.

To distinguish pure bank loan supply movements from other variations, such as demand variations for credit, I consider two empirical measures of bank lending supply. The first is the bank loans share—the share of corporate debt issuance financed via bank loans. This measure is very likely subject to the endogeneity described above.

The second measure is based on bank lending relationships: it captures the exposure of firms to "unhealthy" banks. Typically, banks lend to a large number of firms. Hence, the decision of a bank to lend is likely to be unrelated to a specific firm's performance. Moreover, banks that experienced financial turmoil did so mainly due to their exposure to risky financial instruments such as toxic mortgage loans. Hence, this measure could be considered as an exogenous event for the particular firm and, hence, less prone to endogeneity.

Empirical Proxies for Bank Lending Supply

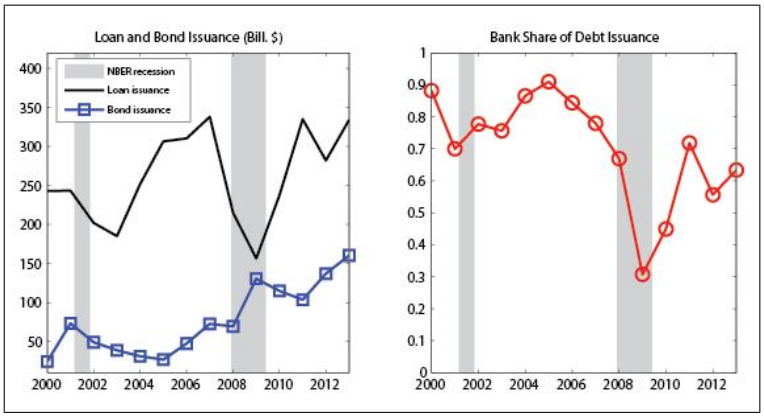
The Bank Share of Debt Issuance

Our first measure of bank lending supply is an aggregate measure: the share of corporate debt issuance financed via bank loans. In particular, we define the bank loan share in period t as

$$\text{Bank Loan Share}_t = \frac{\text{\$ Total Bank Loan Issuance}_t}{\text{\$ Total Debt Issuance}_t}$$

Total debt issuance is defined as the total bank loan issuance plus corporate bond issuance. For corporate bond issuance we use the Securities Data Corporations' New Bond Issuance database, which is again available through Thomson Reuters. Similar to loan issuance, we have information on the amount, issue date, maturity, cost, and issuer name

Figure 2 Loan and Bond Issuance



Notes: Left panel shows the total loan and bond issuances in billions of dollars. Right panel shows the bank loan share of debt issuance. All amounts are deflated using the PPI.

for corporate bond issuances. The screening of bond issuance follows similar steps to the ones for loan issuance.

The left panel of Figure 2 plots the aggregate bond issuance alongside aggregate loan issuance. In contrast to bank loan lending, bond issuance increased between 2007–09. Issuance of new bonds totaled around $80 billion in 2007 and went up to $130 billion during the crisis. This was the result of more firms choosing bond issuance as a means of financing. In particular, the annual number of bond deals increased from around 200 to 400 per year. In contrast, given bond issuance, the average amount of issuance decreased (but less than the decrease in the average loan issuance). In particular, the average amount per bond deal decreased from around $350 million to around $300 million. That means that on average firms substituted bank loan financing with corporate debt issuance. This is consistent with the findings of Adrian, Colla, and Shin (2012).

The right panel of Figure 2 plots the bank share of debt issuance. During the period 2002–07 firms financed (on average) nearly 80 percent of their borrowing using bank loans. During the financial crisis, this share decreased dramatically to 30 percent. As mentioned, this was the result of firms assuming less bank loan debt and at the same time partially substituting loan issuance with corporate debt issuance.

The bank share of debt issuance is a traditional measure of aggregate bank lending conditions also used by Kashyap, Stein, and Wilcox (1993). While the latter paper considers only short-term debt (commercial paper), I consider bonds of all maturities.

Bank Lending Relationships

The second measure of bank lending is based on Chodorow-Reich (2014). While the bank share is an aggregate measure (indexed by period t) this measure is firm-specific. In particular, I measure a firm's exposure to banks that experienced reductions in their lending during the crisis. Being exposed to a bank means being in a business relationship with the bank in the form of acquiring a loan.

Disruptions are measured by the difference in a bank's loan issuance before and after the crisis. Some banks exhibited a sharp reduction in their lending while others maintained a constant flow. An extreme example is Lehman Brothers, which went out of business in September 2008. If a firm was borrowing primarily from Lehman Brothers, then this firm experienced a more severe tightening in its borrowing capacity compared to other firms that were borrowing from other institutions.

The key identification assumptions are 1) the continuation of banking relationships are unrelated to the individual firm's performance, and 2) a disruption in bank lending is firm-specific, i.e. it directly affects a small set of firms.

1. Banks' performance and firms' performance. One question is whether a disruption in a bank's lending is caused by a deteriorating performance of a firm doing business with the bank. There are a couple of reasons why we would expect this not to be the case. First, banks lend to a very large number of firms often from different industries. In our sample, the median bank lends to 1,996 different firms. Hence, a particular firm may be too small to affect the banks' balance sheet. Second, in the recent crisis, banks experienced financial problems depending on their exposure to particular assets such as toxic mortgage loans. Hence, the continuation of lending by a particular bank is likely to not be related to an individual firm's performance.

2. Bank shocks as firm-specific shocks. A typical loan is provided by a group of banks (syndicate). One of these banks—the book manager—leads, originates, structures, and runs the books of the deal. The book manager typically provides the largest portion of the loan. It is rare for a deal to include more than one book manager. The main question here is whether firms use different

Table 2 Total Fraction of Firms Borrowing From a Given Number of Banks

Number of Banks	Fraction of Firms	Average Number of Deals
1	76%	1.9
2	16%	4.4
3	5%	7.2
4	1%	9.9

Note: The table calculates the fraction of firms borrowing from a given number of banks for the period 2000–13. The table also reports how many loan deals have these firms made.

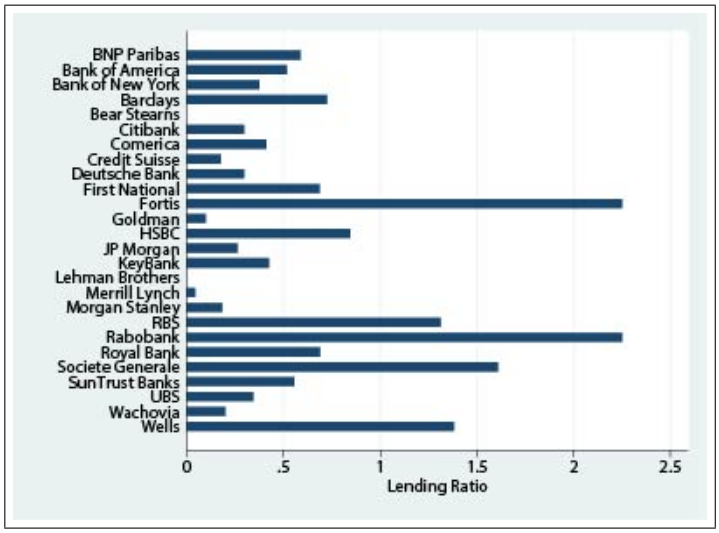
banks for different deals or use the same set of banks for all their deals. In Table 2, I calculate the number of banks that a typical firm uses for borrowing. In our sample, 76 percent of firms are borrowing from just one bank. These firms have made, on average, 1.9 deals, which means there are many firms applying to the same bank for a new loan. Sixteen percent of firms are using two banks for an average of 4.4 deals. Finally, 5 percent and 1 percent of firms are using three and four banks, respectively. These numbers corroborate the hypothesis that firms typically borrow repeatedly from the same set of banks. Hence, it may be natural to think of a bank's performance as a "firm-specific" shock.

The following section describes the construction of our empirical measure for bank lending. First, I calculate how many loans a bank made before and after the crisis. A loan deal is associated with a bank if the bank's name appears as a primary writer of the deal. I perform this calculation for the period October 2005 to June 2007 and the period October 2008 to June 2009. Then for every bank j, I calculate the ratio:

$$\text{Bank Lending Ratio}_j = \frac{18}{8} \times \frac{\#\text{ Loans given by bank } j \text{ in October 2008 - June 2009}}{\#\text{ Loans given by bank } j \text{ between October 2005 - June 2007}} \quad (1)$$

The ratio is multiplied by $\frac{18}{8}$ to adjust for the fact that the numerator accounts for a shorter period (in months) than the denominator. Figure 3 plots the bank lending ratio for a selected group of banks. The median lending ratio is 0.55: after 2008, the median bank gave almost half as many loans as it gave before the crisis. However, there

Figure 3 Bank Lending Ratio



Notes: The figure plots the bank lending ratio: the number of loan deals issued by a bank during October 2008 to June 2009 to the number of loan deals issued by the same bank during October 2005 to June 2007

is a lot of heterogeneity in the lending ratio, with some banks performing much better than others. Lehman Brothers did not give any loans in the period October 2008 to June 2009, so its lending ratio is 0 and the same holds for Bear Stearns. In contrast, institutions such as Wells Fargo, Societe Generale, Rabobank, and Fortis experienced strong lending growth even after the crisis.[2]

The next step is to construct a firm-specific measure of exposure to "unhealthy" banks. To do so, we calculate how much a firm borrowed from a particular bank over the entire sample period 2000–13. We define the weight as

$$w_{i,j} = \frac{\text{\$ Borrowed by firm } i \text{ through bank } j}{\text{Total \$ Borrowed by firm } i}$$

We then define the exposure measure as

[2] The growth of Wells Fargo does not reflect its acquisition of Wachovia in October 2008 since in our data Wachovia exhibits positive growth in loan issuance even after October 2008.

$$DL_i = \sum_j w_{i,j} \times \text{Bank Lending Ratio}_j$$

DL_i summarizes the change in borrowing opportunities by firm i before and after the crisis. If a firm is borrowing heavily from a bank with a low lending ratio, then its borrowing opportunities decreased during the recession and vice versa. If a firm used a balanced borrowing strategy, it is more likely to have a DL_i close to the average lending ratio. It turns out that the average firm has an exposure measure equal to 0.40 with a standard deviation of 0.38.

3. EMPIRICAL SPECIFICATIONS AND RESULTS

In section 2.3, we defined supply-side disruptions to bank lending using two measures: 1) the aggregate bank share and 2) a firm's exposure to "unhealthy" banks. I have argued so far that the second measure is less prone to endogeneity than the first measure. The purpose of this section is to explore how bank lending affects firm-level investment using both measures.

There is a vast literature on the determinants of investment. The prototype paper of Fazzari, Hubbard, and Petersen (1988) tested whether investment depends solely on Tobin's Q or if a firm's cash flows matters as well. Our empirical specification builds on their framework but also includes our variable of interest: bank lending.

In particular, the first specification is

$$\begin{aligned}(\frac{\Delta I}{K})_{i,t} = \beta_0 + \beta_1 (\frac{\text{Cash Flow}}{K})_{i,t} + \beta_2 \log(Q)_{i,t} + \beta_3 \text{ Bank Share }_t + \\ + \ \beta_4 \ \text{Bank Share}_t \times (\frac{\text{Cash Flow}}{K})_{i,t} + \mathbf{X}'_{i,t}\gamma + \varepsilon_{i,t} \qquad (2)\end{aligned}$$

Equation (2) uses the "aggregate bank share" as a measure of bank lending. In this specification, we make use of the panel dimension of our data between 2000–13. Hence, we have information for every firm i at year t. We drop firms that are in our sample for less than four years or firms that do not appear in all consecutive years. The dependent variable $\frac{\Delta I}{K}$ for firm i in period t is the change in investment for firm i between year t and $t-1$ normalized by the firm's total assets in year $t-1$.

As mentioned, we control for the firm's cash flows and Tobin's Q in period t. Tobin's Q for firm i in period t is defined as the firm's common shares outstanding multiplied by the stock price at closing time in period t divided by firm's assets in period t. The main regressor of

interest is "Bank Share," our proxy for bank lending in this specification. Note that bank lending is an average over firms for every period, so it is only indexed by t. We also control for other firm characteristics. In particular, $X'_{i,t}$ is a vector including log-assets, the leverage ratio (debt-to-assets ratio), and a dummy variable indicating whether the firm paid some cash dividends during the year. Also note that this specification allows the inclusion of fixed effects.

In the second specification, the main regressors are a firm's "exposure to unhealthy banks," which serves as a proxy for access to borrowing. The "exposure" measure is firm-specific and is constructed using a ratio over years. Hence, the specification relies on cross-sectional variation (variables only indexed by i but not t). So we cannot include fixed effects here. The regression is

$$\left(\frac{\Delta I_i}{K_{i,2006}}\right) = \beta_0 + \beta_1 \left(\frac{\text{Cash Flow}}{K}\right)_{i,2006} + \beta_2 \log(Q)_{i,2006} + \beta_3 DL_i + \\ + \beta_4 DL_i \times \left(\frac{\text{Cash Flow}}{K}\right)_{i,2006} + \mathbf{X}'_{i,2006}\gamma + \varepsilon_i \tag{3}$$

The dependent variable in equation (3) is defined as

$$\Delta I_i = \frac{\text{Average Investment between 2009 - 2010}}{\text{Average Investment between 2006 - 2008}} \tag{4}$$

Since investment is affected with a lag, we compare investment between 2006–08 to 2009–10. We divide this ratio by assets in our base year 2006. In our specification, we also include the cash flow ratio, Tobin's Q, and covariates for the base year 2006.

For convenience we present in Table 3 the coefficients from a simple regression of bank lending (using both measures) to investment without any controls. The main takeaway is that results change sharply when we switch from one bank lending measure to the other. In the first specification ("aggregate bank share") bank lending is highly procyclical and significant. When the aggregate bank share decreases by 1 percentage point, investment (normalized by assets) decreases by 6.6 percentage points. In contrast, in the second specification ("firm's exposure") the coefficient on bank lending is significant.

As mentioned, Tables 4-12 in the Appendix provide the full set of coefficients for both regressions. In all specifications that include the aggregate bank share, bank lending is strongly correlated with the change in investment. The coefficient is statistically significant and varies between [0.056-0.066]. Consistent with the results of Fazzari, Hubbard, and Petersen (1988), cash flow is an important determinant of investment alongside Tobin's Q. However, when we include the aggregate bank share, cash flow loses its significance.

Table 3 Investment and Bank Lending

	Dependent variable = Change in investment	
Bank Lending Measure	Aggregate Bank Share	Firm's Exposure
Specification	Equation (2)	Equation (3)
Bank Lending	0.066***	-0.001
	(0.003)	(0.000)

Notes: One, two, or three stars represent significance at 1 percent, 5 percent, and 10 percent, respectively.

There does not seem to be any interaction between cash flows and changes in the bank share for the whole sample. However, when we divide the sample between firms with and without access to the bond market (Tables 6-9), surprisingly, the interaction becomes significant for firms with access to bond markets. Moreover, when fixed effects are included (Table 5), size (as proxied by log-assets) is positively related with the change in investment and leverage is negatively related. Dividend payout is negatively related, albeit less statistically significant.

Results from regression 3 are presented in Tables 10-12 in the Appendix. In all specifications the firm's exposure to unhealthy banks is not significant. However, in this specification, the interaction between cash flow and bank supply is positive, which seems to go against the intuition that high-cash-flow firms must be less affected by changes in borrowing opportunities.

4. CONCLUSION

In this article, I examine if bank lending matters for corporate investment. Following Chodorow-Reich (2014), I use DealScan data to construct a firm-specific measure of bank lending supply: the relative exposure of each firm to banks that faced severe lending disruptions. I find that bank lending does not significantly affect investment. In contrast, a traditional measure of bank lending, such as the aggregate bank share of debt issuance, points to a strong relation between bank lending investment.

The exercise focuses on large, publicly listed firms from Compustat. These firms can typically substitute more easily bank lending with other financing tools such as external and internal equity. Hence, it would

be useful for one to use the same methodology to examine the effect of bank lending on small firms. Unfortunately, to my knowledge, data on the investment decision of small firms is not readily available. Hence, we leave this as a future research question.

REFERENCES

Adrian, Tobias, Paolo Colla, and Hyun Song Shin. 2012. "Which Financial Frictions? Parsing the Evidence from the Financial Crisis of 2007–09." Federal Reserve Bank of New York Staff Reports 528 (June).

Becker, Bo, and Victoria Ivashina. 2014. "Cyclicality of Credit Supply: Firm-level Evidence." *Journal of Monetary Economics* 62 (March): 76–93.

Berger, Allen N., and Gregory F. Udell. 1995. "Relationship Lending and Lines of Credit in Small Firm Finance." *Journal of Business* 68 (July): 351–81.

Bernanke, Ben S., and Alan S. Blinder. 1988. "Credit, Money, and Aggregate Demand." *American Economic Review Papers and Proceedings* 78 (May): 435–39.

Chodorow-Reich, Gabriel. 2014. "The Employment Effects of Credit Market Disruptions: Firm-level Evidence from the 2008-9 Financial Crisis." *Quarterly Journal of Economics* 129 (February): 1–59.

Fazzari, Steven M., R. Glenn Hubbard, and Bruce C. Petersen. 1988. "Financing Constraints and Corporate Investment." *Brookings Papers on Economic Activity* 1988 (1): 141–95.

Ivashina, Victoria, and David Scharfstein. 2010. "Bank Lending During the Financial Crisis of 2008." *Journal of Financial Economics* 97 (September): 319–38.

Jiminez, Gabriel, Atif Mian, Jose-Luis Peydro, and Jesus Saurina. 2014. "The Real Effects of the Bank Lending Channel." Working Paper (November).

Kashyap, Anil K., Jeremy C. Stein, and David W. Wilcox. 1993. "Monetary Policy and Credit Conditions: Evidence from the Composition of External Finance." *American Economic Review* 83 (March), 78–98.

Ramey, Valerie. 1993. "How Important is the Credit Channel in the Transmission of Monetary Policy?" *Carnegie-Rochester Conference Series on Public Policy* 39 (December): 1–45.

Table 4 Investment and Bank Lending: Bank Debt Share

Cash Flow/Assets	0.005***	0.006	0.006	0.006	0.007
	(0.001)	(0.006)	(0.006)	(0.006)	(0.006)
log (Q)	0.014***	0.013***	0.013***	0.014***	0.014***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Bank Share		0.056***	0.056***	0.056***	0.056***
		(0.003)	(0.003)	(0.003)	(0.003)
Bank Share		-0.001	-0.001	-0.001	-0.002
x Cash Flow/Assets		(0.009)	(0.009)	(0.009)	(0.009)
Log Assets			0.0002	0.0000	0.0003
			(0.0002)	(0.0002)	(0.0002)
Leverage				0.006***	0.005***
				(0.002)	(0.002)
Dividend Payout					-0.002***
					(0.0008)
Observations	23106	23106	23106	23106	23106
Fixed Effects	No	No	No	No	No
Access to Bond Market	Yes/No	Yes/No	Yes/No	Yes/No	Yes/No

Table 5 Investment and Bank Lending: Bank Debt Share

Cash Flow/Assets	0.005**	0.014	0.019*	0.018*	0.018*
	(0.002)	(0.009)	(0.009)	(0.006)	(0.009)
log (Q)	0.029***	0.026***	0.027***	0.027***	0.027***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Bank Share		0.057***	0.061***	0.062***	0.062***
		(0.003)	(0.003)	(0.003)	(0.003)
Bank Share		-0.012	-0.001	-0.014	-0.014
x Cash Flow/Assets		(0.012)	(0.012)	(0.012)	(0.012)
Log Assets			0.007***	0.007***	0.007***
			(0.001)	(0.001)	(0.001)
Leverage				-0.009**	-0.009**
				(0.004)	(0.004)
Dividend Payout					0.003**
					(0.001)
Observations	23106	23106	23106	23106	23106
Fixed Effects	Yes	Yes	Yes	Yes	Yes
Access to Bond Market	Yes/No	Yes/No	Yes/No	Yes/No	Yes/No

Table 6 Investment and Bank Lending: Bank Debt Share

Cash Flow/Assets	0.005***	0.005	0.006	0.006	0.006
	(0.001)	(0.007)	(0.007)	(0.007)	(0.007)
log (Q)	0.016***	0.015***	0.015***	0.016***	0.015***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Bank Share		0.052***	0.052***	0.052***	0.052***
		(0.004)	(0.004)	(0.004)	(0.004)
Bank Share		-0.000	-0.000	-0.001	-0.001
x Cash Flow/Assets		(0.009)	(0.009)	(0.009)	(0.009)
Log Assets			0.0004	0.0002	0.0007
			(0.0003)	(0.0003)	(0.0003)
Leverage				0.007**	0.006**
				(0.002)	(0.002)
Divident Payout					-0.004***
					(0.001)
Observations	15994	15944	15944	15944	15944
Fixed Effects	No	No	No	No	No
Access to Bond Market	No	No	No	No	No

Table 7 Investment and Bank Lending: Bank Debt Share

Cash Flow/Assets	0.004*	0.001	0.016*	0.015	0.015
	(0.002)	(0.009)	(0.009)	(0.009)	(0.009)
log (Q)	0.030***	0.027***	0.028***	0.027***	0.027***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Bank Share		0.054***	0.058***	0.058***	0.058***
		(0.004)	(0.004)	(0.004)	(0.004)
Bank Share		-0.011	-0.012	-0.012	-0.012
x Cash Flow/Assets		(0.012)	(0.012)	(0.012)	(0.012)
Log Assets			0.006***	0.006***	0.006***
			(0.001)	(0.001)	(0.001)
Leverage				-0.009	-0.008
				(0.005)	(0.005)
Dividend Payout					0.002
					(0.002)
Observations	15944	15944	15944	15944	15944
Fixed Effects	Yes	Yes	Yes	Yes	Yes
Access to Bond Market	No	No	No	No	No

Table 8 Investment and Bank Lending: Bank Debt Share

Cash Flow/Assets	0.012***	-0.264**	-0.264**	-0.264**	-0.264**
	(0.012)	(0.106)	(0.105)	(0.105)	(0.106)
log (Q)	0.009***	0.007***	0.007***	0.007***	0.007***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Bank Share		0.056***	0.057***	0.057***	0.057***
		(0.006)	(0.006)	(0.006)	(0.006)
Bank Share		0.540***	0.553***	0.555***	0.555***
x Cash Flow/Assets		(0.143)	(0.144)	(0.144)	(0.144)
Log Assets			0.001**	0.001**	0.001**
			(0.000)	(0.000)	(0.000)
Leverage				0.004	0.004
				(0.004)	(0.004)
Dividend Payout					0.0003
					(0.001)
Observations	7162	7162	7162	7162	7162
Fixed Effects	No	No	No	No	No
Access to Bond Market	Yes	Yes	Yes	Yes	Yes

Table 9 Investment and Bank Lending: Bank Debt Share

Cash Flow/Assets	0.070	-0.114	-0.102	-0.102	-0.103
	(0.053)	(0.184)	(0.197)	(0.197)	(0.195)
log (Q)	0.026***	0.022***	0.027***	0.027***	0.027***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)
Bank Share		0.060***	0.066***	0.066***	0.066***
		(0.006)	(0.006)	(0.006)	(0.006)
Bank Share		0.252	0.295	0.293	0.294
x Cash Flow/Assets		(0.285)	(0.291)	(0.292)	(0.288)
Log Assets			0.013***	0.013***	0.013***
			(0.002)	(0.002)	(0.002)
Leverage				-0.005	-0.004
				(0.009)	(0.009)
Dividend Payout					0.005
					(0.002)
Observations	15944	15944	15944	15944	15944
Fixed Effects	Yes	Yes	Yes	Yes	Yes
Access to Bond Market	Yes	Yes	Yes	Yes	Yes

Table 10 Investment and Bank Lending: Exposure Measure

Cash Flow/Assets	0.086***	0.063***	0.040***	0.040***	0.040***
	(0.004)	(0.005)	(0.006)	(0.006)	(0.006)
log (Q)	-0.001***	-0.001***	-0.001**	-0.012**	-0.012**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
DL		-0.001	-0.001*	-0.001*	-0.001*
		(0.000)	(0.000)	(0.000)	(0.000)
DL x Cash Flow/ Assets		0.050***	0.053***	0.053***	0.052***
		(0.009)	(0.009)	(0.009)	(0.009)
Log Assets			-0.001***	-0.001***	-0.001***
			(0.000)	(0.000)	(0.000)
Leverage				0.0001	0.0001
				(0.001)	(0.001)
Dividend Payout					0.0005
					(0.0003)
Observations	819	819	819	819	819
Access to Bond Market	Yes/No	Yes/No	Yes/No	Yes/No	Yes/No

Table 11 Investment and Bank Lending: Exposure Measure

Cash Flow/Assets	0.087***	0.057***	0.016	0.016	0.017
	(0.007)	(0.010)	(0.012)	(0.012)	(0.012)
log (Q)	-0.003***	-0.003***	-0.001	-0.0008	-0.0008
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
DL		-0.002*	-0.001*	-0.001	-0.001
		(0.001)	(0.000)	(0.000)	(0.000)
DL x Cash Flow/ Assets		0.059***	0.053***	0.052***	0.052***
		(0.015)	(0.014)	(0.014)	(0.014)
Log Assets			-0.003***	-0.003***	-0.003***
			(0.000)	(0.000)	(0.000)
Leverage				0.002	0.002
				(0.02)	(0.02)
Dividend Payout					0.0001
					(0.0008)
Observations	322	322	322	322	322
Access to Bond Market	No	No	No	No	No

Table 12 Investment and Bank Lending: Exposure Measure

Cash Flow/Assets	0.065***	0.057***	0.034***	0.034***	0.034***
	(0.004)	(0.009)	(0.009)	(0.009)	(0.012)
log (Q)	-0.0002	-0.0003	-0.000	-0.000	-0.000
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
DL		0.000	0.000	0.000	0.000
		(0.0004)	(0.0002)	(0.0002)	(0.0002)
DL		0.026	0.026	0.024***	0.022
x Cash Flow/					
Assets		(0.024)	(0.023)	(0.023)	(0.023)
Log Assets			-0.0006***	-0.0006***	-0.0007***
			(0.000)	(0.000)	(0.000)
Leverage				-0.0007	-0.0007
				(0.0005)	(0.0005)
Dividend Payment					0.0003
					(0.0002)
Observations	497	497	497	497	497
Access to Bond Market	Yes	Yes	Yes	Yes	Yes

Economic Quarterly—Volume 101, Number 4—Fourth Quarter 2015—Pages 323–352

Time-Varying Parameter Vector Autoregressions: Specification, Estimation, and an Application

Thomas A. Lubik and Christian Matthes

Time-varying parameter vector autoregressions (TVP-VARs) have become an increasingly popular tool for analyzing the behavior of macroeconomic time series. TVP-VARs differ from more standard fixed-coefficient VARs in that they allow for coefficients in an otherwise linear VAR model to vary over time following a specified law of motion. In addition, TVP-VARs often include stochastic volatility (SV), which allows for time variation in the variances of the error processes that affect the VAR.

The attractiveness of TVP-VARs is based on the recognition that many, if not most, macroeconomic time series exhibit some form of nonlinearity. For instance, the unemployment rate tends to rise much faster at the start of a recession than it declines at the onset of a recovery. Stock market indices exhibit occasional episodes where volatility, as measured by the variance of stock price movements, rises considerably. As a third example, many aggregate series show a distinct change in behavior in terms of their persistence and their volatility around the early 1980s when the Great Inflation of the 1970s turned into the Great Moderation, behavior that is akin to a structural shift in certain

We are grateful to Pierre-Daniel Sarte, Daniel Tracht, John Weinberg, and Alex Wolman, whose comments greatly improved the exposition of this paper. The views expressed in this paper are those of the authors and not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System. Lubik: Research Department, Federal Reserve Bank of Richmond. P.O. Box 27622, Richmond, VA 23261. Email: thomas.lubik@rich.frb.org. Matthes: Research Department, Federal Reserve Bank of Richmond. P.O. Box 27622, Richmond, VA 23261. Email: christian.matthes@rich.frb.org.

DOI: http://doi.org/10.21144/eq1010403

moments of interest. All these examples of nonlinearity in macroeconomic time series have potentially distinct underlying structural causes. But they can all potentially be captured by means of the flexible framework that is a TVP-VAR with SV.

A VAR is a simple time series model that explains the joint evolution of economic variables through their own lags. A TVP-VAR preserves this structure but in addition models the coefficients as stochastic processes. In the most common application, the maintained assumption is that the coefficients follow random walks, specifically the intercepts, the lag coefficients as well as the variance and covariances of the error terms in the regression. Conditional on the parameters, a TVP-VAR is still a linear VAR, but the overall model is highly nonlinear. While the assumption of random walk behavior may seem restrictive, it provides for a flexible functional form to capture various forms of nonlinearity.

The main challenge in applying TVP-VAR models is how to conduct inference. In this article, we therefore discuss the Bayesian approach to estimating a TVP-VAR with SV.[1] Bayesian inference in this class of models relies on the Gibbs sampler, which is designed to easily compute multivariate densities. The key insight is to break up a computationally intractable problem into sequences of feasible steps. We will discuss these steps in detail and show how they can be applied to TVP-VARs.

The article is structured as follows. We begin with a discussion of the specification of TVP-VARs and how they are developed from fixed-coefficient VARs. We show how to introduce stochastic volatility in the covariance matrix of the errors and present an argument for why time variation in the lag coefficients needs to be modeled jointly with stochastic volatility. The main body of the article presents the Gibbs sampling approach to conducting inference in Bayesian TVP-VARs, which we preamble with a short discussion of the thinking behind Bayesian methods. Finally, we illustrate the method by means of a simple application to data on inflation, unemployment, and the nominal interest rate for the United States.

1. SPECIFICATION

VARs are arguably the most important empirical tool for applied macroeconomists. They were introduced to the economics literature by Sims (1980) as a response to the then-prevailing large-scale macroeconometric modeling approach. What Sims memorably criticized were the

[1] Nakajima (2011) and Doh and Connolly (2012) provide similar overviews of the TVP-VAR methodology.

incredible identification assumptions imposed in these models that stemmed largely from a lack of sound theoretical economic underpinnings and that hampered structural interpretation of their findings. In contrast, VARs are deceptively simple in that they are designed to simply capture the joint dynamics of economic time series without imposing ad-hoc identification restrictions.

More specifically, a VAR describes the evolution of a vector of n economic variables y_t at time t as a linear function of its own lags up to order L and a vector e_t of unforecastable disturbances:

$$y_t = c_t + \sum_{j=1}^{L} A_j y_{t-j} + e_t. \tag{1}$$

It is convenient to assume that the error term e_t is Gaussian with mean 0 and covariance matrix Ω_e. c_t is a vector of deterministic components, possibly including time trends, while the A_j are conformable matrices that capture lag dynamics.

VAR models along the lines of (1) have proven to be remarkably popular for studying, for instance, the effects and implementation of monetary policy (see Christiano, Eichenbaum, and Evans 1999, for a comprehensive survey). However, VARs of this kind can only describe economic behavior that is approximately linear and does not exhibit substantial variation over time. The linear VAR in (1) contains a built-in notion of time invariance: conditional forecasts as of time t, such as $E_t y_{t+1}$, only depend on the last L values of the vector of observables but are otherwise independent of time. More strongly, the conditional one-step-ahead variance is fully independent of time: $E_t[(y_{t+1} - E_t y_{t+1})(y_{t+1} - E_t y_{t+1})'] = \Omega_e$.

Yet, in contrast, a long line of research documents that conditional higher moments can vary over time, starting with the seminal ARCH model of Engle (1982). Moreover, research in macroeconomics, such as Lubik and Schorfheide (2004), has shown that monetary policy rules can change over time and can therefore introduce nonlinearities, such as breaks or shifts, into aggregate economic time series.[2] The first observation has motivated Uhlig (1997) to introduce time variation in Ω_e. The second observation stimulated the work by Cogley and Sargent (2002) to introduce time variation in A_j and c in addition to stochastic volatility.

[2] This feature would make a linear model less suited to capture the true dynamics of the economy. Whether and to what extent linear approximations can be used to analyze environments with time-varying parameters has been studied by Canova, Ferroni, and Matthes (2015).

We will now describe how to model time variation in each of these sets of parameters separately. In the next step, we will discuss why researchers should model changes in both sets of parameters jointly. We then present the Gibbs sampling algorithm that is used for Bayesian inference in this class of models and which allows for easy combination of the approaches because of its modular nature.

A VAR with Random-Walk Time Variation in the Coefficients

Suppose a researcher wants to capture time variation in the data by using a parsimonious yet flexible model as in the VAR (1). The key question is how to model this time variation in the coefficients A_j and c. One possibility is to impose a priori break points at specific dates. Alternatively, break points can be chosen endogenously as part of the estimation algorithm. Threshold VARs or VARs with Markov switching in the parameters (e.g., Sims and Zha 2006) are examples of this type of model, which is often useful in environments where the economic modeler may have some a priori information or beliefs about the underlying source of time variation, such as discrete changes in the behavior of the monetary authority. In general, however, a flexible framework with random time variation seems preferable for a wide range of nonlinear behavior in the data. Following Cogley and Sargent (2002), a substantial part of the literature has consequently opted for a flexible specification that can accommodate a large number of patterns of time variation.

The standard model of time variation in the coefficients starts with the VAR (1). In contrast to the fixed-coefficient version, the parameters of the intercept and of the lag coefficient matrix are allowed to vary over time in a prescribed manner. We thus specify the TVP-VAR:

$$y_t = c_t + \sum_{j=1}^{L} A_{j,t}\, y_{t-j} + e_t. \tag{2}$$

It is convenient to collect the values of the lagged variables in a matrix and define $X_t' \equiv I \otimes (1, y_{t-1}', ..., y_{t-L}')$, where '$\otimes$' denotes the Kronecker product. We also define θ_t to collect the VAR's time-varying coefficients in vectorized form, that is, $\theta_t \equiv vec([c_t\ A_{1,t}\ A_{2,t}\ ...\ A_{L,t}]')$. This allows us to rewrite (2) in the following form:

$$y_t = X_t'\theta_t + e_t. \tag{3}$$

The commonly assumed law of motion for θ_t is a random walk:

$$\theta_t = \theta_{t-1} + u_t, \tag{4}$$

where $u_t \sim \mathcal{N}(0, Q)$ and is assumed to be independent of e_t. A random-walk specification is parsimonious in that it can capture a large number of patterns without introducing additional parameters that need to be estimated.[3] This assumption is mainly one of convenience for reasons of parsimony and flexibility as (4) is rarely interpreted as the underlying data-generating process for the question at hand, but it can approximate it arbitrarily well (see Canova, Ferroni, and Matthes 2015).

Introducing Stochastic Volatility

A second source of time variation in time series can stem from variation in second or higher moments of the error terms. Stochastic volatility, or, specifically, time variation in variances and covariances, can be introduced into a model in a number of ways. Much of the recent literature on stochastic volatility in macroeconomics has chosen to follow the work of Kim, Shephard, and Chib (1998). It is built on a flexible model for volatility that uses an unobserved components approach.[4]

We start from the observation that we can always decompose a covariance matrix Ω_e as follows:

$$\Omega_e = \Lambda^{-1}\Sigma\Sigma'\left(\Lambda^{-1}\right)'. \tag{5}$$

Λ is a lower triangular matrix with ones on the main diagonal, while Σ is a diagonal matrix. Intuitively, the diagonal matrix $\Sigma\Sigma'$ collects the independent innovation variances, while the triangular matrix Λ^{-1} collects the loadings of the innovations onto the VAR error term e, and thereby the covariation among the shocks. It has proven to be convenient to parameterize time variation in Ω_e directly by making the free elements of Λ and Σ vary over time. While this decomposition is general, once priors on the elements of Σ and Λ are imposed, the ordering of variables in the VAR matters for the estimation of the reduced-form parameters, which stands in contrast to the standard time-invariant VAR model (see Primiceri 2005).

We now define the element of Λ_t in row i and column j as λ_t^{ij} and a representative free element j of the time-varying coefficient matrix Σ_t as σ_t^j. It has become the convention in the literature to model the

[3] Different specifications for the time-varying lag coefficients are entirely plausible. For instance, a stationary VAR(1) representation, such as $\theta_t = \overline{\theta} + B\theta_{t-1} + u_t$, can easily be accommodated using the estimation algorithms described in this article.

[4] The approach to modeling stochastic volatility outlined here is the most common in the literature on TVP-VARs, but there are alternatives such as Rondina (2013). Moreover, stochastic volatility models of the form used here are more flexible than ARCH models in that they do not directly link the estimated level of the volatility to realizations of the error process that is being captured.

coefficients σ_t^j as geometric random walks:

$$\log \sigma_t^j = \log \sigma_{t-1}^j + \eta_t^j. \tag{6}$$

For future reference, we collect the σ_t^j in a vector $\sigma_t = \left[\sigma_t^1, ..., \sigma_t^n\right]'$ and the η_t^j in $\eta_t = \left[\eta_t^1, ..., \eta_t^n\right]$, with $\eta_t^n \sim \mathcal{N}(0, W)$ and W diagonal. Similarly, we assume that the nonzero and nonunity elements of the matrix Λ_t, which we collect in the vector $\lambda_t = \left[\lambda_t^{21}, ..., \lambda_t^{n,n-1}\right]$, evolve as random walks:

$$\lambda_t = \lambda_{t-1} + \zeta_t, \tag{7}$$

where $\zeta_t \sim \mathcal{N}(0, S)$ and S block-diagonal.

The error term e_t in the TVP-VAR representation (3) can thus be decomposed into:

$$e_t = \Lambda_t^{-1} \Sigma_t \varepsilon_t, \tag{8}$$

which implicitly defines ε_t. It is convenient to normalize the variance of ε_t to unity. It is thus assumed that the error terms in each of the equations of the model are independent. In more compact form, we can write:

$$V = Var\left[\begin{pmatrix} \varepsilon_t \\ \zeta_t \\ \eta_t \end{pmatrix}\right] = \begin{bmatrix} I & 0 & 0 \\ 0 & S & 0 \\ 0 & 0 & W \end{bmatrix}. \tag{9}$$

The TVP-VAR literature tends to impose a block-diagonal structure for V, mainly for reasons of parsimony since the TVP-VAR is already quite heavily parameterized. Allowing for a fully generic correlation structure among different sources of uncertainty would also preclude any structural interpretation of the innovations. Following Primiceri (2005), the literature has therefore adopted a block-diagonal structure for S, which implies that the nonzero and non-one elements of Λ_t that belong to different rows evolve independently. Moreover, this assumption simplifies inference substantially since it allows Kalman smoothing on the nonzero and non-one elements of Λ_t equation by equation, as we will discuss further below.

Why We Want to Model Time Variation in Volatilities *and* Parameters

A TVP-VAR with stochastic volatility is a heavily parameterized object. While it offers flexibility to capture a wide range of time variation and nonlinear features of the data, it also makes estimation and inference quite complicated. In practice, modelers restrict the covariance matrix of the innovations to the laws of motion for the time-varying

coefficients in order to sharpen inference. Moreover, Bayesian priors are often used to aid inference. Given a need to impose some structure to aid inference, this naturally raises the question whether a TVP-VAR with stochastic volatility is not overparameterized.

One answer to this question relies on the idea that a TVP-VAR can be regarded as the reduced-form representation of an underlying Dynamic Stochastic General Equilibrium (DSGE) model, in which there is time variation. This time variation in the underlying data-generating process (DGP) carries over to its reduced form, which might be, or is approximated by, a TVP-VAR.[5] More specifically, changes, discrete or continuous, in structural parameters carry over to changes in lagged reduced-form coefficients and parameters of the covariance matrix.[6] Hence, a TVP-VAR specification should a priori include stochastic volatility to be able to represent an underlying DSGE model.

A second response is essentially a corollary to the previous point. Sims (2002) argues that a model with only time variation in parameters could mistakenly result in a substantial amount of time variation even though the true DGP only features stochastic volatility. This insight can be illustrated by means of the following simple example, which also shows that the reverse can hold: a modeler could mistakenly estimate stochastic volatility even though the true DGP only features time variation in coefficients.

Consider a univariate AR(1)-process with stochastic volatility:

$$z_t = \rho z_{t-1} + \sigma_t \varepsilon_t, \tag{10}$$

where $|\rho| < 1$, $\varepsilon_t \sim \mathcal{N}(0,1)$, and σ_t is a generic stochastic volatility term, such as the one described above. Suppose an econometrician has access to a sample of data from this DGP, but does not know the true form of the underlying model. In order to investigate the time variation in the data, he proposes a model with only time-varying coefficients instead of stochastic volatility. As a simple rewriting of equation (10) suggests, he could indeed find evidence for time variation in the parameters:

$$z_t = \rho_t z_{t-1} + \tilde{\sigma} \varepsilon_t, \tag{11}$$

[5] It is well-known that in some cases a linear VAR is an exact representation of the reduced form of a DSGE model (see Fernandez-Villaverde et. al. 2007). It is less well-known to what extent this is true for TVP-VARs. For instance, Cogley, Sbordone, and Matthes (2015) show that DSGE models with learning have a TVP-VAR as reduced form.

[6] This insight underlies Benati and Surico's (2009) critique of Sims and Zha's (2006) Markov-switching VAR approach to identifying monetary policy shifts and also Lubik and Surico's (2010) critique of standard empirical tests of the validity of the Lucas critique.

where $\rho_t = \rho + \frac{(\sigma_t - \tilde{\sigma})\varepsilon_t}{z_{t-1}}$.

If the DGP is instead of the form:

$$z_t = \rho_t z_{t-1} + \sigma\varepsilon_t, \tag{12}$$

and estimates a stochastic volatility model on data generated from this model, he would erroneously find evidence of stochastic volatility:

$$z_t = \tilde{\rho} z_{t-1} + \sigma_t \varepsilon_t, \tag{13}$$

where $\sigma_t = \sigma + \frac{(\rho_t - \tilde{\rho})z_{t-1}}{\varepsilon_t}$. Including time variation jointly in coefficients and stochastic volatility therefore allows economists to let the data speak on which of the two sources are more important.

2. ESTIMATION AND INFERENCE

A TVP-VAR with stochastic volatility is a deceptively simple object on the surface, as it superficially shares the structure of standard linear VARs. Estimation and inference in the latter case is well-established and straightforward. Since a linear VAR is a seemingly unrelated regression (SUR) model, it can be efficiently estimated equation by equation using ordinary least squares (OLS). Conducting inference on transformations of the original VAR coefficients, such as impulse response functions, is somewhat more involved yet well-understood in the literature. Estimation and inference in a TVP-VAR, however, reaches a different level of complexity since the model is fundamentally nonlinear due to the time variation in the coefficients and in the covariance matrix of the error terms.

We now describe in detail the standard approach to inference in TVP-VARs. It relies on Bayesian estimation, the basic concepts of which we introduce briefly in the following. Bayesian estimation and inference is conducted using the Gibbs sampling approach, which we go on to discuss at some length. Finally, we discuss how researchers can report and interpret the results from TVP-VAR models in a transparent and efficient manner.

Why a Bayesian Approach?

The standard approach to estimating and conducting inference in TVP-VARs uses Bayesian methodology. The key advantage over frequentist methods is that it allows researchers to use powerful computational algorithms that are particularly well-adapted to the treatment of time variation. Moreover, the use of prior information in a Bayesian framework helps researchers to discipline the behavior of the model, which

is especially relevant in high-dimensional problems such as those discussed in this article.[7]

Bayesian and frequentist inference are fundamentally different approaches to describing and making assessments about data and empirical models. Bayesian inference starts by postulating a prior distribution for the parameters of the model. This prior is updated using the information contained in the data, which is extracted using a likelihood function. The object of interest in Bayesian estimation is the posterior distribution, which results from this updating process. Estimators in a Bayesian context are thus defined as statistics of this distribution such as the mean or mode.

We can describe these basic principles in a somewhat more compact and technical form. Suppose that a Bayesian econometrician is interested in characterizing his beliefs about parameters of interest Θ after having observed a sample of data y^T of length T. The econometrician holds beliefs prior to observing the data, which can be described by the *prior* $p(\Theta)$. Moreover, he can summarize the data by computing the *likelihood function* $p(y^T|\Theta)$, which describes how likely the observed data are for any possible parameter vector Θ. The beliefs held by the econometrician after seeing the data are summarized by the *posterior distribution* $p(\Theta|y^T)$. The relationship between those three densities is given by *Bayes' law*:

$$p(\Theta|y^T) = \frac{p(y^T|\Theta)p(\Theta)}{\int p(y^T|\Theta)p(\Theta)d\Theta}, \tag{14}$$

which describes how to optimally update the beliefs contained in $p(\Theta)$ using data summarized by $p(y^T|\Theta)$. The posterior $p(\Theta|y^T)$ is a distribution on account of normalization by the *marginal data density* $\int p(y^T|\Theta)p(\Theta)d\Theta$, which is the joint distribution of data y^T and parameters Θ after integrating out Θ. It can serve as a measure of fit in this Bayesian context.

Bayesian estimation ultimately consists of computing the posterior distribution. Bayesian inference rests on the moments of this distribution. It does not require any arguments about limiting behavior as $T \to \infty$, since from a Bayesian perspective y^T is fixed and is all that is needed to conduct inference. On the other hand, the challenges for Bayesian econometricians are virtually all computational in that: (i) the likelihood function has to be evaluated; (ii) the joint distribution of prior and likelihood has to be computed; and (somewhat less

[7] This is not to say that frequentist inference does not introduce prior information by, for instance, imposing bounds on the parameter space. The use of Bayesian priors, however, makes this more explicit and generally more transparent.

crucially) (iii) the marginal data density has to be obtained. What aids in this process is the judicious use of priors and fast and robust methods for characterizing $p(y^T|\Theta)p(\Theta)$. This can be accomplished in Bayesian VARs by means of the Gibbs sampler.

Gibbs Sampling of a TVP-VAR

Characterizing a posterior distribution is a daunting task. Except in special cases, analytical solutions for given prior and likelihood densities are not available. Conducting inference via describing the posterior with its moments is thus not an option. As evidenced by the seminal textbook of Zellner (1971), much of Bayesian analysis before the advent of readily available computing power and techniques was concerned with finding conjugate priors for a large variety of problems. A conjugate prior is such that when confronted with the likelihood function, the posterior distribution is of the same family as the prior. However, as a general matter this path proved not to be a viable option as many standard Bayesian econometric models do not easily yield to analytical characterization.

This changed with the development of sampling and simulation methods that allow researchers to characterize the shape of an unknown distribution. These methods are built on the idea that when a large sample from a known density is available, sample moments approximate population moments very well by the laws of large numbers. Consequently, Bayesian statisticians have developed methods to efficiently sample from unknown posterior densities indirectly by sampling from known densities. Once the thus-generated sample is at hand, sampling moments can be used to characterize the posterior distribution.[8]

The basic idea behind the Gibbs sampler is to split the parameters Θ of a given model into b blocks Θ^1,Θ^2,...,Θ^b.[9] The Gibbs sampler proposes to generate a sample from $p(\Theta|y^T)$ by iteratively sampling from $p(\Theta^j|y^T, \Theta^{-j})$, $\forall j = 1, ..., b$, where Θ^{-j} denotes the entire parameter vector except for the jth block. This approach rests on the idea that the entire set of conditional distributions fully characterizes the joint distribution under fairly general conditions. At first glance, nothing

[8] The exposition here is intentionally, but unavoidably, superficial. Readers interested in the technical issues underlying the arguments we make here are referred to some of the excellent textbooks on Bayesian inference such as Robert and Casella (2004) or Gelman et al. (2014).

[9] Generally, there are no restrictions placed on the relative size of the blocks. In fact, the blocking scheme, that is, its individual size, could be random. However, for time-varying parameter models, one particular blocking scheme turns out to be especially useful.

much has been gained: we have broken up one large inference problem into a sequence of smaller inference problems, namely characterizing the conditional distributions $p(\Theta^j|y^T, \Theta^{-j})$ instead of the full distribution. In the end, there is no guarantee that this makes the inference problem more tractable.

However, Bayesian statisticians have developed closed forms for posterior distributions for some special cases. The ingenuity of the Gibbs sampler is thus to break up a large intractable inference problem into smaller blocks that can then be evaluated independently and sequentially. The challenge is to find a blocking scheme, a partition of the set of parameters, that admits closed-form solutions for the posteriors conditional on all other parameters of the model. In the case of TVP-VARs, such blocking schemes have been developed by Cogley and Sargent (2002), Primiceri (2005), and Del Negro and Primiceri (2015).[10]

A Motivating Example for the Gibbs Sampler

In order to illustrate the basic idea behind Gibbs sampling, we consider a simple fixed-coefficient AR(1) model:

$$z_t = \rho z_{t-1} + \sigma \varepsilon_t, \tag{15}$$

where $\varepsilon_t \sim \mathcal{N}(0,1)$. The parameters of interest are ρ and σ, on which we want to conduct inference. The first step in deriving the Gibbs sampler is to specify priors for these parameters. We assume the following priors:

$$\rho \sim \mathcal{N}(\mu_\rho, V_\rho), \tag{16}$$

$$\sigma^2 \sim \mathcal{IG}(a, b), \tag{17}$$

where IG denotes the inverse Gamma distribution with scale and location parameters a and b, respectively.

The likelihood for this standard AR(1) model is given by $L(\rho, \sigma) = p(z_0) \prod_{t=1^T} p(z_t|z_{t-1})$, which is written as the product of conditional distributions $p(z_t|z_{t-1})$ and the likelihood of the initial observation $p(z_0)$. As is common practice, we drop the term $p(z_0)$ and instead work with the likelihood function $L(\rho, \sigma) = \prod_{t=1}^{T} p(z_t|z_{t-1})$. Defining $Y = [z_1\ z_2\ z_3 \dots z_T]'$ and $X = [z_0\ z_1\ z_2 \dots z_{T-1}]'$, the likelihood is given by:

$$L(\rho, \sigma) = (2\pi)^{-T/2} (\sigma^2)^{-T/2} \exp\left[-\frac{1}{2\sigma^2} (Y - X\rho)'(Y - X\rho) \right]. \tag{18}$$

[10] Computer code to estimate this class of models is available from Gary Koop and Dimitris Korobilis at: http://personal.strath.ac.uk/gary.koop/bayes_matlab_code_by_koop_and_korobilis.html

Combining this expression with the priors listed above using Bayes' Law gives the joint posterior of ρ and σ^2, conditional on the data:

$$p\left(\rho,\sigma|Y,X\right) \propto L(\rho,\sigma)\times$$
$$\exp\left[-\frac{1}{2}\left(\rho-\mu_\rho\right)' V_\rho^{-1}\left(\rho-\mu_\rho\right)\right]\times\left(\sigma^2\right)^{-[a+1]}\exp\left(-\frac{1}{b\sigma^2}\right), \quad (19)$$

where the first term is the likelihood function, the second is the prior on the autoregressive coefficient ρ, and the third term is the prior on the innovation variance σ^2. Although we can identify and compute analytically the individual components of the posterior, the posterior distribution for $\rho,\sigma|Y,X$ is unknown.

The Gibbs sampler allows us to partition the parameter set into separate blocks for ρ and σ, for which we can derive the conditional distributions. After some algebra, we can find the conditional posterior distributions:

$$\rho|\sigma,Y,X \sim \mathcal{N}\left[\begin{array}{c}\left(X'X/\sigma^2+V_\rho^{-1}\right)^{-1}\left(X'Y/\sigma^2+V_\rho^{-1}\mu_\rho\right),\\ \left(X'X/\sigma^2+V_\rho^{-1}\right)^{-1}\end{array}\right], \quad (20)$$

$$\sigma^2|\rho,Y,X \sim \mathcal{IG}\left[T/2+a, b^{-1}+\frac{1}{2}(Y-X\rho)'(Y-X\rho)\right]. \quad (21)$$

The conditional posteriors for ρ and σ have known distributions, which can be sampled by using standard software packages. The procedure would be to start with an initial value for σ^2 and then draw from the conditional distribution $\rho|\sigma,Y,X$. Given a draw for ρ, in the next step we would sample from the conditional distribution $\sigma^2|\rho,Y,X$. Repeated iterative sampling in this manner results in the joint posterior distribution $\rho,\sigma|Y,X$.

The Gibbs sampler can be applied to models with time-varying parameters in a similar manner, the key step being the application of a blocking scheme for which the conditional distributions are either known or from which it is easy to generate samples. The additional challenge that TVP-VARs present is that the parameters of interest are not fixed coefficients, but are themselves time-series processes that are a priori unobservable. The general approach to dealing with unobservable components is the application of the Kalman filter if the model can be cast in a state-space form. In the following, we discuss how these two additional techniques can be used to estimate TVP-VARs.

Linear Gaussian State-Space Systems

Bayesian estimation relies on the ability of the researcher to cast a model in a form such that it is amenable for sampling. The Gibbs

sampler provides one such technique. A second crucial component of inference in a TVP-VAR is the state-space representation, which connects variables that are observed, or are in principle observable, to those that are unobserved. Conceptually, Bayesian estimation produces a time series and its density for the time-varying components of the TVP-VAR by means of the Kalman filter as applied to a linear Gaussian state-space system. This specification has the advantage that the posterior distribution is known analytically for a Gaussian prior on the initial state.

Specifically, a state-space system can be defined as follows:

$$y_t = A_t x_t + B_t v_t, \tag{22}$$

$$x_t = C x_{t-1} + D w_t, \tag{23}$$

where y_t denotes a vector of observables and x_t a vector of possibly unobserved states. v_t and w_t are Gaussian innovations, each element of which is independent of the others with mean 0 and variance 1. A_t, B_t, C, and D are *known* conformable matrices. The standard approach for deriving the posterior for x_t in this system was developed by Carter and Kohn (1994), which builds on the Kalman filter and which we discuss in the next section.

Application of the Kalman filter to a state-space system allows the modeler to construct a sequence of Gaussian distributions for $x_t|y^t$, that is, the distribution of the unobservable state x at time t, conditional on the observables y^t, where a superscript denotes the entire sample up to that point.[11] As it turns out, various blocks of the Gibbs sampler for a TVP-VAR model take the form of linear Gaussian state-space systems. The challenge is to find blocks for the parameters in the TVP-VAR such that each block fits this Gaussian state-space structure. The fundamental nonlinearity of the TVP-VAR can thus be broken up into parts that are conditionally linear and from which it can be easily sampled. As long as each block has a tractable structure conditional on other blocks of parameters, the Gibbs sampler can handle highly nonlinear problems.

The Kalman Filter

The Kalman filter is a widely used method for computing the time paths of unobserved variables from a Gaussian state-space system. We now briefly review and present the equations used for drawing a sequence of the unobserved states (conditional on the entire set of observations

[11] If the modeler is instead interested in the distributions $x_t|y^T$, where T denotes the sample size, the Carter-Kohn algorithm draws paths of the unobserved state variable x_t for $t = 1, ..., T$ conditional on the entire sample of observables y^T.

$y_1, ..., y_T$). A more detailed discussion and explanation can be found in Primiceri (2005).

The system is assumed to take the form (22)-(23). We want to draw from the distribution $p(x_1, ..., x_T|y_1, ..., y_T)$.[12] It can be shown that $p(x_1, ..., x_T|y_1, ..., y_T) = p(x_T|y_T) \prod_{t=1}^{T} p(x_t|x_{t+1}, y_1, ..., y_t)$. To generate draws from each of these densities, we first run the Kalman filter to calculate the mean and variance of the state x_t conditional on data up to time t. We assume a prior for x_0 that is Gaussian with mean $x_{0|0}$ and variance $V_{0|0}$. The Kalman filter is then summarized by the following equations:

$$x_{t|t-1} = Cx_{t-1|t-1} \tag{24}$$

$$V_{t|t-1} = CV_{t-1|t-1}C' + DD' \tag{25}$$

$$K_t = V_{t|t-1}A_t' \left(A_t V_{t|t-1} A_t' + B_t B_t'^{-1}\right) \tag{26}$$

$$x_{t|t} = x_{t|t-1} + K_t(y_t - A_t x_{t|t-1}) \tag{27}$$

$$V_{t|t} = V_{t|t-1} - K_t A_t V_{t|t-1} \tag{28}$$

These equations produce $x_{t|t} = E(x_t|y_1, ..., y_t)$ and the associated conditional variance $V_{t|t}$. The conditional distributions of the states are Gaussian.

We can generate a draw for $x_T|y_1, ..., y_T$ by using the conditional mean and variance for period T. Once we have such a draw, we can recursively draw the other states (x_{t+1} denotes a draw of the state for period $t + 1$):

$$x_{t|t+1} = x_{t|t} + V_{t|t}CV_{t+1|t}^{-1}(x_{t+1} - Cx_{t|t}) \tag{29}$$

$$V_{t|t+1} = V_{t|t} - V_{t|t}C'V_{t+1|t}^{-1}CV_{t|t} \tag{30}$$

In the following, we will now discuss each Gibbs sampler step in turn, which builds on the Kalman filter.

The Choice of Priors

The first step in Bayesian analysis is to choose the priors on the parameters of the model. In contrast to a frequentist approach, the model parameters in a Bayesian setting are random variables. Since a Gibbs sampler proceeds iteratively, we impose priors on the initial values of the TVP-VAR parameters. Conceptually, it is therefore useful to distinguish between two sets of parameters: the parameters associated with the coefficients and innovation terms in the representation (4)

[12] We do not explicitly state the dependence of the densities in this section on the system matrices A, B, C_t, and D_t, but as we show later this can be handled by the right conditioning and sequencing within the Gibbs sampler.

and the parameters governing the law of motion of the time-varying terms. More specifically, we impose priors on $(\theta_0, \Lambda_0, \log \Sigma_0)$ and on (Q, W, S), respectively.

The initial values of the lag coefficient matrices θ_0, of the free elements of the loading matrix in the innovation terms Λ_0, and of the independent innovation variances $\log \Sigma_0$ are assumed to have normally distributed priors:

$$\theta_0 \sim \mathcal{N}(\overline{\theta}, \kappa_\theta V_\theta), \tag{31}$$

$$\Lambda_0 \sim \mathcal{N}(\overline{\Lambda}, \kappa_\Lambda V_\Lambda), \tag{32}$$

$$\log \Sigma_0 \sim \mathcal{N}(\overline{\Sigma}, I), \tag{33}$$

where $\overline{\theta}$, $\overline{\Lambda}$, and $\overline{\Sigma}$ are the prior means of the respective variables, while V_θ and V_Λ are their prior covariance matrices. The covariance matrix of the prior on $\log \Sigma_0$ is normalized at unity. κ_θ and κ_Λ are scaling parameters that determine the tightness of the priors.

We also have to choose priors for the covariance matrices of the innovations in the law of motions for the above-referenced TVP-VAR parameters. These are, respectively, the innovation variance for the lag coefficient matrices, Q; for the error variance, W; and for the loading matrix, S. As is common for covariance matrices in Bayesian analysis, the priors follow an Inverted Wishart distribution:

$$Q \sim \mathcal{IW}(\kappa_Q^2 df_Q V_Q, df_Q), \tag{34}$$

$$W \sim \mathcal{IW}(\kappa_W^2 df_W V_W, df_W), \tag{35}$$

$$S \sim \mathcal{IW}(\kappa_S^2 df_S V_S, df_S), \tag{36}$$

where κ are the scaling factors, df the degrees of freedom, and the matrices V the respective variances.

A key issue is how to choose the parameters for the priors. Cogley and Sargent (2005) and Primiceri (2005) propose using a constant-coefficient VAR estimated on a training sample to initialize the prior means and the matrices V. The coefficients $\left(\overline{\theta}, \overline{\Lambda}, \overline{\Sigma}\right)$ and (V_θ, V_Λ) can then be directly computed from a least-squares regression. Nevertheless, this still leaves substantial degrees of freedom as there is no clear guideline on how to choose the training sample. The scaling parameters κ turn out to be important as they govern the prior amount of time variation. Primiceri (2005) estimates the κ on a small grid of values using a time-consuming reversible-jump MCMC algorithm that, as a preliminary step, requires estimation of the model for each possible combination of parameters. Following Primiceri, most researchers

have chosen to use his estimated values regardless of the application at hand.[13]

The Ordering of Blocks in a TVP-VAR

Once the priors have been chosen, the next step involves combining the prior distribution with the likelihood function. In a Bayesian approach, the resulting posterior distribution contains all information that is available to the researcher, which includes the prior and the observed data as encapsulated in the likelihood. Moreover, and in contrast to a frequentist approach to inference, Bayesian estimation does not involve an actual estimation step, where an estimator, that is, a mapping from data to the object of interest that satisfies some desirable criteria, is derived. Bayesian estimation simply involves characterizing the posterior distribution, which can be accomplished in the case of a TVP-VAR by means of the Gibbs sampler. A Bayesian econometrician then finds it often convenient to report moments of the posterior as estimation results.

The Gibbs sampler relies on the idea that it is often much easier to sequentially sample from conditional distributions, whose probability laws may be known, than from an unknown distribution. The tricky and often difficult part of this approach is to partition the parameter space into blocks such that this sampling is feasible and can be accomplished efficiently. To wit, in the full TVP-VAR model with both time-varying parameters and stochastic volatility, we need to estimate the following set of parameters: θ^T, Σ^T, Λ^T, Q, S, and W, where the T superscripts indicate that there can be in general sample size T parameters.

In the following, we describe the Gibbs sampler proposed by Del Negro and Primiceri (2015), which is based on the original contribution of Primiceri (2005). As a matter of notation, we also introduce a set of auxiliary variables s^T that are used for the estimation of the stochastic volatilities. In subsequent sections we discuss the drawing of each of those blocks in more detail. Even more detailed descriptions can be found in Primiceri (2005) or Koop and Korobilis (2010).

Conceptually, the two main steps of the Gibbs sampler involve drawing the covariance matrix of the independent innovations in the TVP-VAR, Σ^T, conditional on the data, the other coefficient vectors,

[13] In the recent literature, there has been much interest in the role that these scaling parameters play, in particular the hyperparameters for Q, S, and W. As it turns out, choice of these parameters can affect estimation results along many dimensions. For a recent application that studies the importance of these hyperparameters in producing the 'correct' inference see Lubik, Matthes, and Owens (2016).

and the covariance matrices of the processes governing time variation. In the second step, the remaining parameters are drawn from a distribution conditional on the data and on the draw from the first step Σ^T. Specifically, the procedure is to

1. draw Σ^T from $p(\Sigma^T|y^T, Q, S, W, \Lambda^T, \theta^T, s^T)$

2. draw Λ^T, θ^T, s^T, Q, S, and W from $p(Q, S, W, \Lambda^T, \theta^T, s^T|y^T, \Sigma^T)$.

The second step is implemented as a sequence of intermediate steps. First, the algorithm draws from $p(Q, S, W, \Lambda^T, \theta^T|y^T, \Sigma^T)$, while the auxiliary variables s^T are then drawn from $p(s^T|Q, S, W, \Lambda^T, \theta^T, y^T, \Sigma^T)$. This second step is split up into these two parts since this blocking scheme allows drawing θ^T without having to condition on s^T. Specifically, the sequence is to

i) draw Λ^T from $p(\Lambda^T|y^T, \Sigma^T, Q, S, W, \theta^T)$

ii) draw Q, S and W from $p(Q, S, W|y^T, \Lambda^T, \Sigma^T, \theta^T)$

iii) draw θ^T from $p(\theta^T|y^T, Q, S, W, \Lambda^T, \Sigma^T)$

iv) draw s^T from $p(s^T|y^T, \theta^T, Q, S, W, \Lambda^T, \Sigma^T)$.

Drawing Σ^T

The first step of the Gibbs sampler involves generating draws of the elements of covariance matrix Σ^T from a distribution that is conditional on the data y^T and the remaining coefficient matrices. This conditional distribution conflates elements of the prior and the likelihood function; it is, in fact, a marginal density of the posterior. Draws are realizations of the random variable Σ^T and are accordingly recorded. We now describe how a known conditional probability distribution for Σ^T can be derived under this blocking scheme.

We can rewrite equation (3) under the assumption that e_t features stochastic volatility:

$$\Lambda_t \left(y_t - X_t'\theta_t\right) = y_t^* = \Sigma_t \varepsilon_t, \tag{37}$$

where we have made use of the decomposition of the errors in equation (8). Given the conditioning set of this block in Step 1 above, y_t^* is known. We can nowcast this representation into a Gaussian state-space system to draw the elements of Σ^T. Squaring each element of this vector and taking natural logarithms yields for each element i of

y_t^*:[14]

$$\log\left((y_{i,t}^*)^2\right) = y_{i,t}^{**}. \tag{38}$$

We define σ_t as the vector of the diagonal elements of Σ_t.

We then get the state-space system:

$$\begin{aligned} y_t^{**} &= 2\log(\sigma_t) + 2\log(\varepsilon_t), & (39)\\ \log(\sigma_t) &= \log(\sigma_{t-1}) + \eta_t. & (40) \end{aligned}$$

This is a linear state-space system with y_t^{**} being the observable variable, while $\log(\sigma_t)$ is the unobserved state variable. However, it is not Gaussian: each element of $2\log(\varepsilon_t)$ is distributed as $\log\chi^2$ since it is the log of the square of a standard-normal random variable. These shocks can be approximated with a mixture of seven normal variables, as suggested by Kim, Shephard, and Chib (1998). In this step, the auxiliary variables s^T are introduced to provide a record of which of the seven mixture components is 'active' for each element of $2\log(\varepsilon_t)$. Given this approximation, we have another Gaussian state-space system, which can now be evaluated using the Kalman filter. The prediction formulas listed above can be used to generate realizations, that is, draws, of the unobservable σ_t over time.

Drawing Λ^T

Given the draws for the matrix Σ^T, which is a component of the reduced-form error matrix $\Omega_{e,t}$ per equation (8), we can now sample its other component, namely the loadings Λ^T. The first step is to rewrite equation (3) but utilizing a different blocking:

$$\Lambda_t(y_t - X_t'\theta_t) = \Lambda_t\hat{y}_t = \Sigma_t\varepsilon_t. \tag{41}$$

The difference to the previous sampling scheme for Σ^T is that we now condition on Σ^T and are interested in sampling the free elements of the lower-triangular matrix Λ_t.

We can therefore rewrite the equation above by moving elements of $\Lambda_t\hat{y}_t$ to the right-hand side. We can write:

$$\hat{y}_t = Z_t\lambda_t + \Sigma_t\varepsilon_t, \tag{42}$$

where Z_t is a selection matrix that contains elements of the vector $\hat{y}_t$. Together with the set of equations (7), this equation forms a linear Gaussian state-space system. The fact that Z_t depends on elements of $\hat{y}_t$ poses no problem for the sampling step under the assumption

[14] In practice, and in order to improve numerical stability, we instead define $\log((y_{i,t}^*)^2 + c) = y_{i,t}^{**}$, where c is a small 'offset' constant.

that the innovation covariance matrix for λ, S, is block diagonal. The Kalman filter can then be used to obtain draws for Λ^T.

Drawing Innovation Covariance Matrices

In the next step, we are drawing from the innovation covariance matrices for the processes governing the time variation of the VAR parameters. As discussed above, each of the matrices Q, S, and W is assumed to have an inverse-Wishart prior to facilitate the application of the Kalman filter within a Gaussian state-space system. In combination with a normally distributed likelihood, this prior forms a conjugate family since the innovations in the laws of motion for parameters and volatilities are Gaussian. Consequently, the posterior will also be of the inverse-Wishart form, which has a closed-form representation.[15] It is then straightforward to sample the innovation covariance matrices by drawing from the known inverted-Wishart posterior.

Drawing θ^T

In a penultimate step, we are now ready to sample from the conditional distribution for the TVP-VAR coefficient matrices. Given the preliminary work up this point and the use of the conditioning scheme that we describe above, this is now straightforward. Since we condition on draws for the covariance matrix of e_t, which in the general model with stochastic volatility will consist of draws for Λ_t and Σ_t, equations (3) and (4) form a Gaussian state-space system. We can sample from the posterior distribution for θ^T in the manner described above by using the Kalman prediction equations to sequentially construct the draws.

Drawing s^T

The final step that brings everything together involves the auxiliary variables s^T that we use to track the stochastic volatilities. As we discuss above, each element of s_t is drawn from a discrete distribution, a mixture of normals, with seven possible outcomes. Denote the prior probability for outcome j as q_j. The conditional posterior probability used to drawing outcome j for each element of s^T is then proportional to

$$q_j f_N(y_{i_t}^{**}, 2\log(\sigma_{i,t}) + m_j, v_j^2), \tag{43}$$

[15] See, for example, Gelman et al. (2014).

where m_j and v_j are the given mean and standard deviation of each element of the Gaussian approximation and $f_N(x, a, b)$ is the Gaussian density with argument x, mean a, and variance b.

Reporting the Results

Estimating a Bayesian TVP-VAR is tantamount to sampling from a posterior distribution. While the posterior summarizes all information available in the data and in the prior, it is an unwieldy object in that it is a multivariate distribution of which only the conditional distributions are known. The Gibbs sampling algorithm solves this problem by sequentially building up the joint distribution from the conditional distributions. Yet, what Bayesian estimation delivers are distributions and not point estimates. Reporting the results in a manner that is useful for economic interpretation therefore requires some thought. The Bayesian literature focuses on posterior means or medians as counterparts to frequentist point estimates. Instead of standard errors and confidence intervals, Bayesians report coverage regions that essentially are regions of the posterior distribution in which a given percentage of draws fall around a focal point such as the mean or the median.

The results from Bayesian fixed-coefficient VARs can be reported in a similar manner as for frequentist approaches. The reporting problem is compounded, however, in the case of TVP-VARs, since the distribution of the VAR parameters potentially changes at every data point, which is the very definition of time variation. Instead of reporting a single distribution in the case of a fixed-coefficient VAR, the Bayesian econometrician now faces the challenge of reporting a sequence of distributions. We describe in the following how to approach this issue for the case of impulse response functions, which are key objects in the toolkits of time series econometricians.

Impulse Responses

VARs can be used to study the effects of exogenous shocks, that is, of unpredictable changes in the economy. For this purpose, the main tool in VAR analysis is the impulse response function that describes the behavior of a variable in response to a shock over time. In order to understand the sources of business cycles or to analyze policy, it is often desirable to give these shocks a structural interpretation. By doing so, researchers can link the shocks to economic theories.[16] However, the

[16] In line with time-invariant VARs, the literature usually focuses on studying the effects of shocks to observables, not shocks to the parameters that vary over time.

shocks that are estimated as residuals from a regression of the type (1) are generally not useful for this purpose as they conflate the effects of underlying structural disturbances. That is, the estimated residuals are generally correlated, in which case it is not possible to identify the effects of an individual disturbance.

More specifically, a researcher may be interested in uncovering uncorrelated disturbances w_t that are a linear function of the regression errors e_t:

$$H_t w_t = e_t, \tag{44}$$

where it is assumed that w_t is Gaussian with mean zero and a covariance matrix that is normalized to unity, $w_t \sim \mathcal{N}(0, I)$. The conformable matrix H_t is used to transform the errors e_t into the structural shocks w_t. How to derive and impose restrictions on H_t is one of the key issues in VAR analysis. For instance, the economic theories used to define the shocks, e.g., DSGE models, can be used to derive restrictions on H_t. For the most part, it is common practice in the VAR literature to focus on imposing few enough restrictions so that the restrictions do not alter the likelihood function of the model. This has the advantage that the researcher can first estimate a statistical, 'reduced-form' model without worrying about the restrictions used to derive structural shocks. Structural shocks can then be studied after the estimation step is completed.[17]

For purposes of exposition we now discuss the most common and straightforward method for identifying structural shocks. It only assumes restrictions on the within-period-timing of shocks. The specific idea is that some shocks may be causally prior to other shocks in the sense that they have an impact on some variables and not on others within the period. The easiest way to implement this restriction is to make H_t lower triangular. This can be achieved by calculating the Cholesky decomposition of the covariance matrix of the forecast errors.

In the context of TVP-VARs, this type of recursive ordering is appealing because $\Lambda_t^{-1}\Sigma_t$ already has lower triangular form so that the matrix H_t can be directly calculated from the output of the Gibbs sampler. Given H_t, the impulse responses can then be calculated by simulation.[18] In contrast to fixed-coefficient VARs, it is thus not

[17] Using more restrictions so that the likelihood function is altered relative to the estimation of a reduced-form model means that the restrictions have to be imposed during estimation, that is, a 'structural model' has to be estimated directly. This is not often carried out, even though algorithms are now available even in the context of TVP-VARs, for instance in Canova and Perez-Forero (2015).

[18] A simpler method to approximate impulse responses is to draw a set of parameters from the Gibbs sampler output for each time period t and then compute im-

possible to separate the estimation from the identification stage. In this case, the estimated variance-covariance matrix can be decomposed into its recursive components after the VAR is estimated. A detailed description of the algorithm is available in Canova and Gambetti (2009). We briefly describe the algorithm below.

Conceptually, we can define an impulse response as the difference between the expected path of the variables in the model when a shock of a given size hits and the expected path of the same variables when all shocks are drawn randomly from their distributions. In order to calculate impulse responses starting at time t, the first step is to draw a set of parameters from the Gibbs sampling output. Next, paths of future time-varying parameters and volatilities and a sequence of w shocks are simulated once the identification matrix H_t is computed. These objects are then used to calculate one path for the variables of interest using equation (2). The same exercise is repeated, but with the value of one structural shock fixed at one point in time, leaving all other structural shocks at the simulated values. This yields another path for the variables of interest, so that the difference between the paths is one realization of the impulse response. This sequence is repeated a large number of times for different parameter draws from the posterior and different simulated values of parameter paths and shocks. The approach produces a distribution of a path for the impulse responses for each time period in the sample. To report the results, the literature usually either picks a subset of time periods and then plots the median response as well as posterior bands for each time period separately or authors focus on the posterior median responses and plot those over time and for different horizons in a three-dimensional plot.[19]

3. APPLICATION: A SIMPLE TVP-VAR MODEL FOR THE UNITED STATES

We now apply the methods discussed above to three key economic variables: the inflation rate, the unemployment rate, and a nominal interest rate. These three variables form the core of many models that are used to analyze the effects of monetary policy, such as the standard New Keynesian framework. Moreover, they are staples in most VARs that are used for the analysis of monetary policy. In his seminal paper,

---

pulse responses as if those parameters at time t were parameters of a fixed coefficient VAR. This approach is computationally easier but neglects the fact that parameters and volatilities can change in the future.

[19] An example of the former can be found in Benati and Lubik (2014), while the latter approach is used in Amir-Ahmadi, Matthes, and Wang (2016).

Primiceri (2005) estimates a TVP-VAR in these three variables to study the effects of monetary policy in the post-World War II period in the United States. We base our application on his specification.

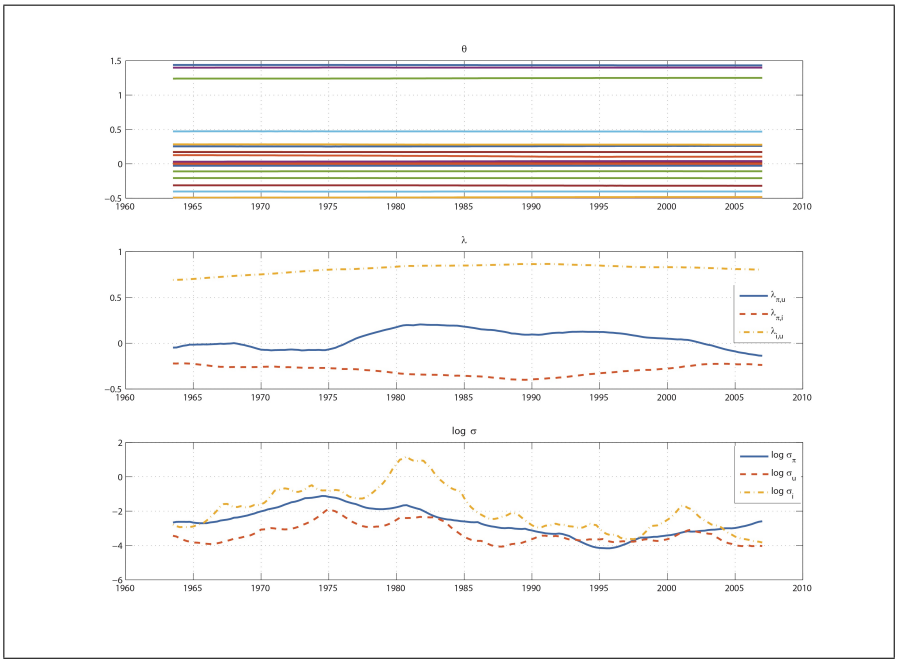
We update the data set to include more recent observations. The full sample ranges from the first quarter of 1953 to the first quarter of 2007, before the onset of the Great Recession. The data are collected quarterly, whereby percentage changes are computed on a year-over-year basis. As our measure of inflation, we use the (log-difference of the) GDP deflator, reported in percentage terms. As our economic activity variable, we pick the headline unemployment rate, while we use the three-month Treasury bill rate as the nominal interest rate variable. The data series are extracted from the FRED database at the Federal Reserve Bank of St. Louis.

We follow Primiceri (2005) in selecting a lag length of two for the TVP-VAR. This choice has become common in the TVP-VAR literature. In fixed-coefficient VARs, a higher number of lags is usually used, but the higher degree of complexity and dimensionality imposes nontrivial computational constraints. A lag length of two thus seems a good compromise and also allows for direct comparison of our results with other key papers in the literature. As discussed above, we need to provide an initialization for the prior. We follow common practice and use the first ten years of data for this purpose. The remaining priors are as in Primiceri (2005).

The first set of results that we extract from our TVP-VAR is contained in Figure 1. We report the median coefficient estimates from our model in three separate panels. The plots start with the first quarter of 1963 because the first ten years of the sample were used for the initialization of the prior. The upper panel contains plots of the time-varying lag coefficients $A_{j,t}$ and the intercept c_t from equation (2). The overriding impression is that there is not much time variation in the lag coefficients. This is a finding that occurs throughout much of the TVP-VAR literature. However, evidence of some more time variation is apparent from the middle and lower panels, which report the time-varying components of the reduced-form innovation variance $\Omega_{e,t} = \Lambda_t^{-1}\Sigma_t\Sigma_t'\left(\Lambda_t^{-1}\right)'$.

The middle panel contains the nonzero and nonunity elements of the lower triangular matrix Λ_t^{-1}. The three off-diagonal elements are thus related to the correlation pattern in the estimated covariance matrix of the shocks. The panel shows that the relationship between inflation and the interest rate errors is consistently negative throughout the sample, while it is positive between the interest rate and unemployment. This observation corresponds to the notion, at least in a reduced-form sense,

Figure 1 Estimated Coefficients



that the interest rate and unemployment move in the same direction while the interest rate and inflation rate move in the opposite direction.

The coefficient $\lambda_{\pi,u}$ for the relationship between inflation and unemployment in the middle panel exhibits more variation. It is positive from 1976 until 2002 and negative before and after. Despite uncertainty surrounding this estimate (not reported), it reveals changes in how unemployment and inflation have interacted over the sample period. This observation is of particular interest since the relationship between these two variables is sometimes described as the Phillips curve, which may embody a trade-off for the conduct of monetary policy. That this trade-off apparently changed in the late 1970s and again in the early 2000s is noteworthy. Finally, the lower panel of Figure 1 depicts the series for the elements of the Σ_t, which is a diagonal matrix. Movements in these terms indicate the extent to which volatility of the estimated errors has changed. The most variation is attributed to the interest rate, followed by the inflation rate.

Figure 1 summarizes all coefficient estimates θ_t from the TVP-VAR with stochastic volatility in a comprehensive manner. The lesson to take away from this is that almost all of the time variation in the

post-World War II history of the three variables appears to be due to stochastic volatility and not to changes in the lag coefficients. This observation is thus conceptually in line with the argument presented in Sims and Zha (2006), who use a Markov-switching VAR and also attribute changes in the behavior of the U.S. business cycle to regime changes in the shocks.

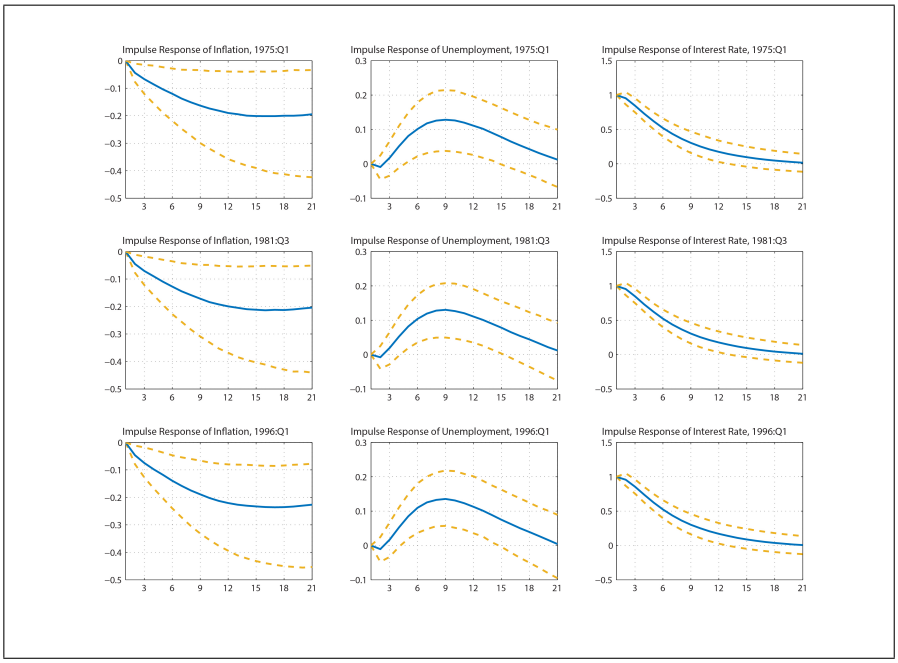
However, we want to raise some caveats for this interpretation. First, the relative importance of variations in the shocks versus changes in the parameters is a long-standing issue in econometrics, ranging from test for structural change (Lubik and Surico 2010) to the proper conditioning of state-space models including unobserved components (Stock and Watson 2003). Disentangling the relative importance of time-variation in the shocks and in lag coefficients is a challenge that a Bayesian approach has not overcome, but the judicious use of priors gives some structure to the issue. Specifically, the choice of an initial prior is informed by a pre-sample analysis, whereby the data stem from the same underlying data-generating process as the latter part of the sample.

Second, there is a concern that TVP-VARs with SV have a tendency to attribute time variation in the data to the stochastic volatility part of the model and not to the lag coefficients. In a simple example above, we argue that the inclusion of stochastic volatility is necessary to avoid a pitfall in the opposite direction. Lubik, Matthes, and Owens (2016) address this aspect in a simulation study based on an underlying nonlinear model and judge that a TVP-VAR does in fact come to the right conclusion as to the sources of time variation, but that a judicious choice of prior is crucial.

The second set of results are reported in Figure 2. These are the impulse response functions of inflation, unemployment, and the interest rate itself to a unit, that is, a 1 percentage point, increase in the three-month nominal rate bond rate. As discussed above, there are impulse responses functions at every single data point, so reporting the full set becomes a challenge. We therefore pick three dates from each decade that are associated with, respectively, the height of a deep recession, the onset of the Volcker disinflation, and at the early stages of a long expansion: 1975:Q1, 1981:Q3, and 1996:Q1. For identification purposes, the variables are in the order: inflation, unemployment, and the interest rate. This implies that the interest rate has no contemporaneous effect on inflation and unemployment, but that it responds contemporaneously to shocks in these two variables. In discussing the result, we focus on the effects of monetary policy shocks.

Figure 2 shows that the impulse responses are remarkably similar across all three time periods. This has already been indicated by the

Figure 2 Impulse Response Functions



observation from Figure 1 that the estimated lag coefficients exhibit virtually no time variation. Since the impulse responses are functions of the lag coefficients, this clearly carries over. The structural responses are also functions of the matrix H_t and therefore related to the factors of the reduced-form error covariance matrix, Λ_t^{-1} and Σ_t, which show more variation; yet, this does not carry over to the impulse responses despite the sign change of the elements of Λ_t^{-1}.

Following a unit innovation, the interest rate returns slowly over time to its long-run level, which it reaches after five years. The response is fairly tightly estimated based on the 90 percent coverage regions. The interest rate's own response in the last column of the figure is very much the same in all periods. On impact, the response of the unemployment rate to a contractionary interest rate shock is zero by construction. Afterward, unemployment starts to rise slowly until hitting a peak around the two-year mark. It returns to its starting value after five years. The unemployment response is much less precisely estimated, with zero included in the coverage region for the first year after impact. Again, the responses across episodes are remarkably similar. An additional point to note is that the median extent of a 1

percentage point interest rate rise is a 0.12 percentage point increase in the unemployment rate. Finally, the interest rate hike reduces inflation over time with a fairly wide coverage region and very similar responses in each of the three time periods.

4. CONCLUSION

This article discusses and reviews the concept and the methodology of time-varying parameter VARs. This class of empirical models has proved to be a flexible and comprehensive approach to capturing the dynamics of macroeconomic series. We focus on the specification and implementation of TVP-VARs in a Bayesian framework since it offers unique computational challenges. To this effect, we present the Gibbs sampler as a convenient and adaptable method for inference. We illustrate the approach by means of a simple example that estimates a small-scale TVP-VAR for the United States.

The TVP-VAR literature is still in its infancy, and there are several issues we plan to address in further detail in a companion article to the present one. Identification of structural shocks is a key element of time-series analysis. The application in the present article uses a simple, yet widely used, recursive identification scheme that is not without its problems. Alternative identification schemes, such as long-run restrictions and sign restrictions, warrant additional consideration although they present unique challenges in a TVP-VAR with SV context. A second issue is to what extent TVP-VARs are able to capture a wide variety of nonlinear behavior in macroeconomic time series, especially when compared to alternative methods, such as regime-switching VARs.

REFERENCES

Amir-Ahmadi, Pooyan, Christian Matthes, and Mu-Chun Wang. 2016. "Drifts and Volatilities under Measurement Error: Assessing Monetary Policy Shocks over the Last Century." *Quantitative Economics,* forthcoming.

Benati, Luca, and Thomas A. Lubik. 2014. "Sales, Inventories, and Real Interest Rates: A Century of Stylized Facts." *Journal of Applied Econometrics* 29 (November/December): 1210–22.

Benati, Luca, and Paolo Surico. 2009. "VAR Analysis and the Great Moderation." *American Economic Review* 99 (September): 1636–52.

Canova, Fabio, and Fernando J. Perez-Forero. 2015. "Estimating Overidentified, Nonrecursive, Time-Varying Coefficients Structural Vector Autoregressions." *Quantitative Economics* 6 (July): 359–84.

Canova, Fabio, Filippo Ferroni, and Christian Matthes. 2015. "Approximating Time Varying Structural Models with Time Invariant Structures." Federal Reserve Bank of Richmond Working Paper 15-10 (September).

Canova, Fabio, and Luca Gambetti. 2009. "Structural Changes in the US Economy: Is There a Role for Monetary Policy?" *Journal of Economic Dynamics and Control* 33 (February): 477–90.

Carter, C. K., and R. Kohn. 1994. "On Gibbs Sampling for State Space Models." *Biometrika* 81 (September): 541–53.

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans. 1999. "Monetary Policy Shocks: What Have We Learned and To What End?" In *Handbook of Macroeconomics, vol. 1*, edited by John B. Taylor and Michael Woodford. North Holland: Elsevier, 65–148.

Cogley, Timothy, and Thomas J. Sargent. 2002. "Evolving Post-World War II U.S. Inflation Dynamics." In *NBER Macroeconomics Annual 2001, vol. 16*, edited by Ben S. Bernanke and Kenneth Rogoff. Cambridge, Mass.: MIT Press, 331–88.

Cogley, Timothy, and Thomas J. Sargent. 2005. "Drift and Volatilities: Monetary Policies and Outcomes in the Post WWII U.S." *Review of Economic Dynamics* 8 (April): 262–302.

Cogley, Timothy, Argia Sbordone, and Christian Matthes. 2015. "Optimized Taylor Rules for Disinflation When Agents are Learning." *Journal of Monetary Economics* 72 (May): 131–47.

Del Negro, Marco, and Giorgio Primiceri. 2015. "Time Varying Structural Vector Autoregressions and Monetary Policy: A Corrigendum." *Review of Economic Studies*, forthcoming.

Doh, Taeyoung, and Michael Connolly. 2012. "The State Space Representation and Estimation of a Time-Varying Parameter VAR with Stochastic Volatility." Federal Reserve Bank of Kansas City Working Paper 12-04 (July).

Engle, Robert F. 1982. "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica* 50 (July), 987–1008.

Gelman, Andrew, et al. 2014. *Bayesian Data Analysis*. Third Edition. Boca Raton: CRC Press.

Fernandez-Villaverde, Jesus, et al. 2007. "ABCs (and Ds) of Understanding VARs." *American Economic Review* 97 (June): 1021–26.

Kim, Sangjoon, Neil Shephard, and Siddhartha Chib. 1998. "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models." *Review of Economic Studies* 65 (July): 361–93.

Koop, Gary, and Demetrios Korobilis. 2010. "Bayesian Multivariate Time Series Methods for Empirical Macroeconomics." Manuscript.

Lubik, Thomas A., Christian Matthes, and Andrew Owens. 2016. "Beveridge Curve Shifts and Time-Varying Parameter VARs." Manuscript.

Lubik, Thomas A., and Frank Schorfheide. 2004. "Testing for Indeterminacy: An Application to U.S. Monetary Policy." *American Economic Review* 94 (March): 190–217.

Lubik, Thomas A., and Paolo Surico. 2010. "The Lucas Critique and the Stability of Empirical Models." *Journal of Applied Econometrics* 25 (January/February): 177–94.

Nakajima, Jouchi. 2011. "Time-Varying Parameter VAR Model with Stochastic Volatility: An Overview of Methodology and Empirical Applications." IMES Bank of Japan Discussion Paper 2011-E-9 (March).

Primiceri, Giorgio E. 2005. "Time Varying Structural Vector Autoregressions and Monetary Policy." *Review of Economic Studies* 72 (July): 821–52.

Robert, Christian, and George Casella. 2004. *Monte Carlo Statistical Methods*. Second Edition. New York: Springer Verlag.

Rondina, Francesca. 2013. "Time Varying SVARs, Parameter Histories, and the Changing Impact of Oil Prices on the US Economy." Manuscript.

Sims, Christopher A. 1980. "Macroeconomics and Reality." *Econometrica* 48 (January): 1–48.

Sims, Christopher A. 2002. "Comment on Cogley and Sargent's 'Evolving Post-World War II U.S. Inflation Dynamics.'" In *NBER Macroeconomics Annual 2001, vol. 16,* edited by Ben S. Bernanke and Kenneth Rogoff. Cambridge, Mass.: MIT Press, 373–79.

Sims, Christopher A., and Tao Zha. 2006. "Were There Regime Switches in U.S. Monetary Policy?" *American Economic Review* 96 (March): 54–81.

Stock, James H., and Mark M. Watson. 2003. "Has the Business Cycle Changed and Why?" In *NBER Macroeconomics Annual 2002, vol. 17*, edited by Mark Gertler and Kenneth Rogoff. Cambridge, Mass.: MIT Press, 159–230.

Uhlig, Harald. 1997. "Bayesian Vector Autoregressions with Stochastic Volatility." *Econometrica* 65 (January): 59–74.

Zellner, Arnold. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York: J. Wiley and Sons, Inc.