When More Is Better: Assessing the Southeastern Economy with Lots of Data

PEDRO SILOS AND DIEGO VILÁN

Silos is a research economist and assistant policy adviser in the macropolicy section, and Vilán is an economist in the regional section, both in the Atlanta Fed's research department. *The authors thank Marco del Negro, John Robertson, and Ellis Tallman for very useful comments.*

Although the economy of each of the six southeastern U.S. states has unique and defining characteristics, these states' business cycles also tend to move together as if responding to some common, underlying factor. Currently, no single economic indicator exists for the economy of the Sixth Federal Reserve District as a whole, which encompasses the entire states of Georgia, Florida, and Alabama and parts of Louisiana, Mississippi, and Tennessee.[1] Rather, analyses of each southeastern state's economy are typically performed individually and then aggregated into a weighted average index. An indicator that captured the overall trend of the region's economy using information from all six states would aid in understanding the unique features of the region's business cycle. These features, when compared to those of the nation or other Federal Reserve districts, could assist in identifying crucial differences and similarities used to develop more accurate forecasts and in turn support monetary policy formulation.

This article outlines and estimates a model that provides such an indicator. We model economic activity in the Sixth District as being driven by an unobserved common factor. Economic activity is measured by a large set of time series of employment, construction, earnings, and sales tax revenues. Disaggregated information for each state is incorporated in a large model from which the common component is derived.

A thorough understanding of the dynamics behind this common factor will enable academics, policymakers, and businesspeople to make a better diagnosis of the condition of the region's economy. In addition, having one comprehensive measure of economic activity for the Sixth District will not only allow for a simplified and faster interpretation of several (sometimes contradicting) economic signals but will also make comparisons with other Federal Reserve district economies easier.

The study also seeks to compare its latent common factor model with the current practice of averaging individual states' coincident indicators. Overall we find that our indicator provides a more reasonable assessment of large idiosyncratic shocks, such as Hurricane Katrina, than the weighted-average estimates. In other words, our model's results provide a better fit to what may be observed a priori in the data, as measured in the aggregate national income and product accounts (NIPA) from the U.S. Bureau of Economic Analysis. Moreover, the indicator provides insights about the different trajectory of the southeastern economy compared to the U.S. economy as a whole.

The Methodology

The model presented in this article is primarily based on the coincident indicator approach pioneered by Stock and Watson (1989) and is closely related to Otrok and Whiteman (1998). In the latter study, the authors develop an indicator that has since been used at the Institute for Economic Research at the University of Iowa to evaluate conditions in the Iowa economy. Because our time series are so lengthy, it is not efficient to apply that study's methodology here. Instead, we follow closely the approach used by Otrok, Silos, and Whiteman (2003), in which the estimation of the unobserved factor is done sequentially rather than in one "block" as in Otrok and Whiteman (1998). This way of sampling avoids some technical difficulties that are related to the sample size. Although our basic setup and idea are the same as Stock and Watson's, our choice of powerful simulation tools allows us to use a large cross-section of series (literally dozens) in contrast to the four or five that Stock and Watson use in their coincident index.[2]

The Setup

With standard assumptions on distributions and functional forms, we construct artificial observations of the common component using a powerful tool called Gibbs sampling (described in more detail in a later section and in the appendix).[3]

We observe n variables, denoted y_{it}, $i = 1,\ldots,n$, that reflect economic activity (employment, income, tax revenues, etc.) during period $t = 1,\ldots,T$. Each i refers to a specific data series; for example, $i = 1$ could be employment in Georgia, while $i = 2$ could be employment in Florida. There is a single common factor, F_t, that accounts for all comovement among the n variables. We assume that this factor is latent (that is, unobserved) and that it can be interpreted as an indicator of the stage of economic conditions or the business cycle in the economy being considered. Clearly, various factors could be affecting the comovement of two or more series, but in this study we are mainly interested in the common factor driving the comovement among all the variables in our data set. We assume that the relationship between any of the series and the common factor is linear:

$$(1)\quad y_{it} = \gamma_i F_t + \varepsilon_{it}.$$

The idiosyncratic dynamics (the dynamics in the individual series that are caused by something other than the common factor) are given by the errors ε_{it}. These error terms follow an autoregressive process:

$$(2)\quad \varepsilon_{it} = \varphi_{i,1}\varepsilon_{i,t-1} + \varphi_{i,t-2} + \nu_{it};\ \nu_{it} \sim N(0,\sigma_i^2).$$

Finally, the equation that governs the dynamics of the common factor has an autoregressive structure as well:

$$(3)\quad F_t = \rho_1 F_{t-1} + \rho_2 F_{t-2} + \omega_t;\ \omega_t \sim N(0,1).$$

Two identification problems arise for the model above. First, the sign for the dynamic factor and the sign of the γ_i are not independently identified. We solve this problem using two normalizations (see the appendix for details).

It should be clear from the above equations that if the common factor F_t were observed, the analysis of this system would be straightforward. In such a textbook case, equations (1) and (2) would form a series of n independent regressions in which errors have an autoregressive structure. The latent factor, however, poses some estimation difficulties. Fortunately, for such difficulties sampling methods developed in the Bayesian statistics literature can be helpful.

The final goal of the estimation is to obtain moments of interest (means, medians, standard deviations, etc.) from a density function (distribution) of the parameters and the unobserved factor given the observed time series data. Bayesian statisticians call this distribution the posterior distribution.[4] Denoting the vector of parameters by θ, the time series by Y, and the unobserved factor by F, let us write this distribution as $p(\theta,F|Y)$.

Sampling from the posterior distribution directly is generally difficult for a large number of time series, each of which is associated with several parameters, while at the same time keeping track of the unobserved factor. Fortunately, Gibbs sampling makes it possible to split this unmanageable distribution into several "sampling blocks." These sampling blocks are themselves density functions but of smaller dimension. The smaller dimension is the result of conditioning on values of parameters that belong to other blocks. For instance, in a very simple setup in which there are no unobserved factors and only two parameters, κ_1 and κ_2, our goal would be to sample from the joint posterior distribution of κ_1 and κ_2 given some data Y, $p(\kappa_1, \kappa_2|Y)$. The Gibbs sampling would allow us to sample sequentially from two conditional posteriors, $p(\kappa_1|Y,\kappa_2)$ and $p(\kappa_2|Y,\kappa_1)$. Of course, in this simple example there seems to be little computational gain from splitting the distribution. However, in problems of large dimension, Gibbs sampling could be the only feasible way of attacking a problem.

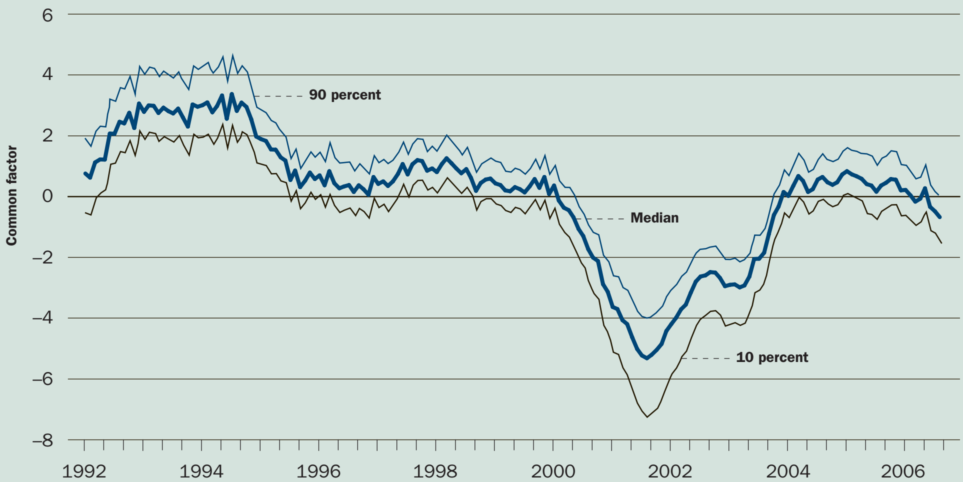
In our application the sampling blocks are as follows: The first is the distribution of the unobserved factor given θ; we have one such distribution for each point in time. The second block is the distribution of ρ_1 and ρ_2 given the unobserved factor. Finally, for each of the $i = 1,\ldots,n$ time series, we would sample γ_i, $\varphi_{i,1}$, $\varphi_{i,2}$, and σ_i^2 given the common factor. In this step we are treating the factor as observed data and therefore dealing with the simple task of obtaining n independent regressions for the "true" observed data.[5] By repeating these steps many times, starting with a guess for the parameter vector θ, the procedure generates a sample for the entire posterior distribution.

---

1. Throughout the article, the terms "Sixth Federal Reserve District" (and shortened forms), "Southeast," and "region" will be used interchangeably.
2. Crone and Clayton-Matthews (2005), using the Stock and Watson methodology, construct individual indicators for each of the fifty states using an approach similar to ours. Here, we estimate the common component jointly for all the states in the Sixth District using a larger data set (for example, we introduce sales tax data).
3. For a clear introduction to Gibbs sampling, see Casella and George (1992).
4. In general, before starting the analysis, the econometrician will combine prior information about the distribution of the unknown parameters, called the prior distribution and denoted by $p(\theta)$, with the "likelihood" of observing the data, given values for the parameters and the unobserved factor. This combination yields the posterior distribution.

   A detailed analysis of how to draw inferences from the posterior is beyond the scope of this article. See the appendix for a more technical overview than the one provided in the text.
5. This mechanism of generating data is, in essence, "data augmentation" (see Tanner and Wong 1987).

Figure 1
Distribution of the Model's Common Factor



Note: Data are from May 1992 through November 2006.

Data Description

Most of the series in this application are not seasonally adjusted. To avoid problems associated with seasonality we run the model in year-over-year growth rates. The only exception is the Georgia Purchasing Managers Index (PMI), which displays no trend, and therefore we estimate it in levels. Moreover, the data are standardized to avoid having to estimate an intercept and because we are mainly interested in comovement among variables.[6] In addition, having all series in a similar scale facilitates the estimation.

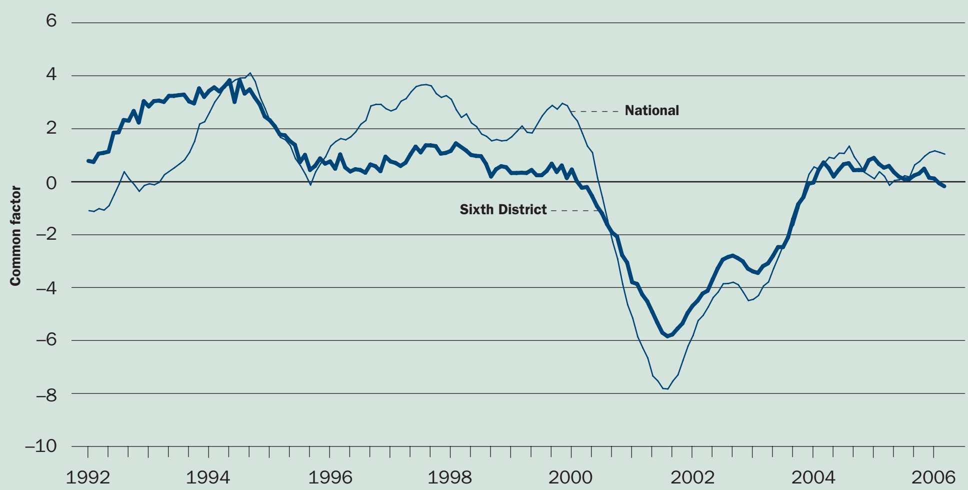
In total, we use twenty-four data series that fall into five groups—nonfarm employment, housing starts, sales tax revenues, average hourly earnings, and Georgia's PMI—described below. Data are monthly, starting in January 1991 to December 2006 (except for hourly earnings, which start in January 2001).

Employment. The employment series includes total nonfarm payroll employment for all six southeastern states. The nonfarm series include payroll data from construction, trade, transportation and utilities, information, financial activities, professional and business services, education and health services, leisure and hospitality, and government sectors. Employment figures are from the U.S. Bureau of Labor Statistics (BLS) and are monthly and seasonally adjusted.

Housing starts. Given that the housing industry accounts for about a quarter of all investment spending and around 5 percent of the overall economy, the housing starts series is considered a leading indicator. The series includes all new privately owned housing units started in each of the six states in the district. Series are seasonally adjusted annual rates from the Bank of Tokyo–Mitsubishi UFJ.

Real hourly earnings. To transform average hourly earnings (AHE) into real terms, we deflate them using the U.S. urban consumer price index (CPI).[7] In this article, earnings are for the manufacturing sector in each state. The BLS data begin in 2001 and are monthly and not seasonally adjusted.

Figure 2
The Model's Common Factor for the Sixth Federal Reserve District and National Economies



Note: Data are from May 1992 through July 2006.

Georgia PMI. The PMI report is a composite index based on five major indicators: new orders, inventory levels, production, supplier deliveries, and employment environment. The Association of Purchasing Managers surveys over 300 purchasing managers nationwide that represent twenty different industries. The PMI data, from Kennesaw State University's Econometric Center, are monthly and not seasonally adjusted.

Sales tax revenues. Sales tax revenues are an important indicator of each state's fiscal strength and, indirectly, of the current regional business cycle conditions. Series data, from Haver Analytics, are monthly and not seasonally adjusted.

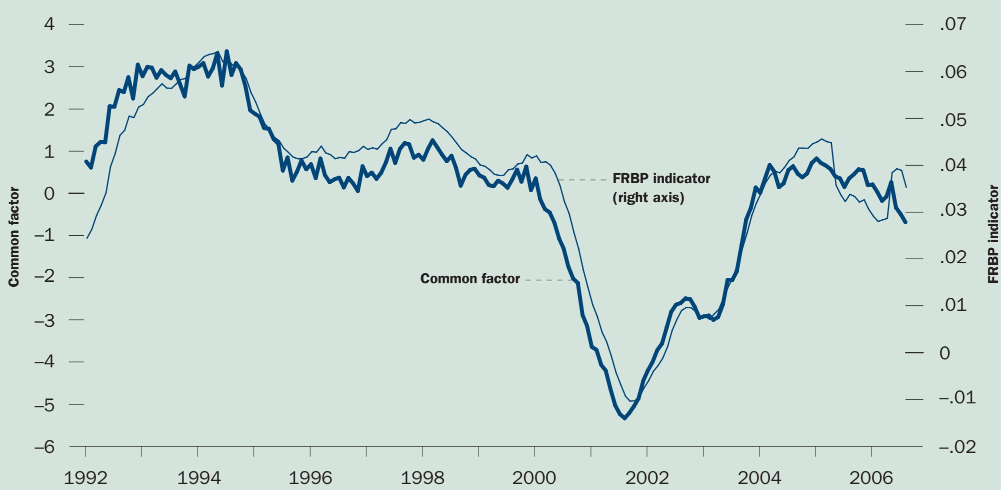
Estimation Results

To summarize the results of our estimation, we first describe the evolution of the unobserved component for the Sixth District and then compare that indicator to an analogous indicator for the U.S. economy. Finally, we compare the estimate obtained here with an indicator of economic activity constructed from series provided by the Federal Reserve Bank of Philadelphia.

Figure 1 shows the median of the common component along with the 10th and 90th percentiles. It is important to realize that the common component is a random variable, and as such it has a distribution at each point in time. The percentiles are plotted along the median to give an idea of the uncertainty of that distribution. At first glance, it is easy to distinguish the recovery from the 1991 recession during the first half of the nineties, the 1994 soft landing caused by the contraction in residential

6. "Standardization" implies that from each series we subtract its mean and then divide by the standard deviation. As a result, all series prior to estimation have a sample mean of zero and a sample standard deviation of one.
7. One can deflate earnings by the Southeast CPI provided by the BLS, but the difference in results is quantitatively insignificant.

Figure 3
The Model's Common Factor and the Federal Reserve Bank of Philadelphia Averaging Indicator for the Six Southeastern States



Note: Data are from May 1992 through December 2006.

investment, and the subsequent slowdown of economic expansion during the second half of the decade. The plunge of economic activity coincides with the recession of 2001, followed by a recovery during the 2003–05 period. One can conclude that the underlying (median) factor reflects the prior notions about the evolution of the region's economy during the past fourteen years.

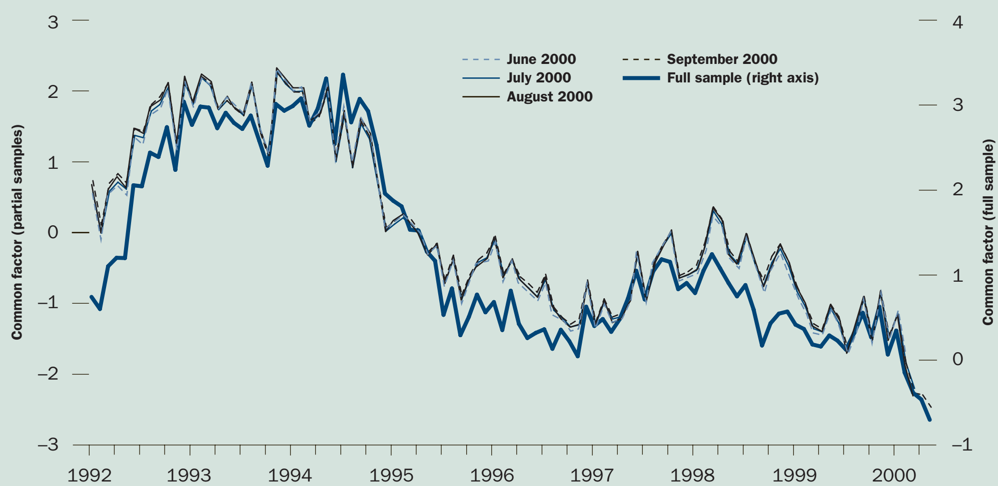
Figure 2 compares the Sixth District economy with the U.S. economy. The graph plots the common factor computed as described above along with the common factor for the national data. The national factor is computed using the same methodology and the series used by the National Bureau of Economic Research (NBER) to date the stages of the business cycle.[8] The figure clearly shows that the two series are highly correlated (the correlation is 0.86). There are, however, a few differences. First, the Sixth District seems to have benefited more than the U.S. economy during the initial recovery after the 1991 recession as illustrated by the concave (downward) shape of the Sixth District common factor compared to the concave (upward) shape of the U.S. factor.

During the dip in economic activity during 1994 and 1995, both series' declines were similar in magnitude. However, although the district economy was able to outgrow the national economy during the recovery in the years after the 1990–91 recession, apparently it did not benefit as much from the boom in the second half of the nineties. Our index reports a much stronger expansion at the national than at the district level.

Additionally, both series fell starting in 2001, although the Sixth District seems to have endured a milder slowdown in economic activity than the overall U.S. economy. Not surprisingly, the recovery following the slowdown was also less pronounced in the district's states than in the rest of the nation. In recent years both series seem to be trailing quite closely.

Figure 3 compares two indicators for the Sixth District: our dynamic factor and the averaging indicators for the six individual southeastern states obtained from the Federal Reserve Bank of Philadelphia (FRBP). The FRBP indicators are constructed with a

Figure 4
Comparison of the Model's Common Factor Estimated Using Different Samples



Note: The full sample uses data from May 1992 through December 2006, but for comparison to the partial samples the graph shows data only through September 2000.

dynamic factor model with four series for each state: employment, hours, the unemployment rate, and real wages. The weights for averaging the six states are given by the states' gross state products (GSPs), and the weighted average is shown in Figure 3.

Comparing our common component and the FRBP indicator, one can see that both series move closely together. There are, nonetheless, two differences between the estimates. The first is related to the impact of Hurricane Katrina on the district's economic activity. While the FRBP index depicts a strong drop in the aftermath of Katrina, our index describes a much smoother trend. We believe, given the relative weight that the state of Louisiana has in the district, that ours is a more accurate estimate.

The second difference in the two indicators appears at the beginning of the plunge in 2001. Our indicator starts falling a few months before the FRBP indicator. This pattern could stem from our use of housing starts data, which are traditionally a leading indicator of the business cycle. Is the early drop an artifact of estimating the model for the entire sample, or would the fall be signalled as well if the data ran only until 2000?[9] To determine whether this leading property of our indicator is a "real time" property, we estimate four additional indicators in which we vary the sample size. The first indicator ends in June 2000, the second in July 2000, the third in August 2000, and the fourth in September 2000. The results of this estimation are plotted in Figure 4, which clearly shows that the magnitude and timing of the fall for a given data point do not depend on the sample size. All four lines lie on top of each other, and the shape of the indicators is very similar to the one obtained when using the full sample.

8. These series are industrial production, real income less transfer payments, employment, and retail sales and inventories.
9. Note that our definition of real time does not take into account data revisions because we assume that the econometrician has the already revised data at the end of her sample.

Conclusion

Evaluating economic conditions generally involves keeping track of literally dozens of time series describing different aspects of an economy. Central bankers, financial institutions, and many corporations and individuals comb through data on labor, products, and factors markets to assess the current state of an economy and make judgments about its future state.

This article applies a methodology to extract a "signal" from a large array of time series representing economic activity and uses that methodology to construct an economic indicator for the Sixth Federal Reserve District. The article outlines the idiosyncrasies of the southeastern economy relative to the U.S. economy and compares the new indicator with a weighted average of indicators for individual states constructed using a similar methodology. The indicator demonstrated here should be of interest to anyone analyzing the condition of the southeastern economy.

Appendix
Constructing the Gibbs Sampler

More detailed expositions of constructing the Gibbs sampler are available in Kim and Nelson (1999) and Otrok, Silos, and Whiteman (2003), but this appendix provides a broad idea of how the algorithm is structured.

Recall that the original model was:

$$y_{it} = \gamma_i F_t + \varepsilon_{it};$$
$$\varepsilon_{it} = \varphi_{i,1}\varepsilon_{i,t-1} + \varphi_{i,t-2} + \nu_{it};\ \nu_{it} \sim N(0,\sigma_i^2);\text{ and}$$
$$F_t = \rho_1 F_{t-1} + \rho_2 F_{t-2} + \omega_t;\ \omega_t \sim N(0,1).$$

As mentioned in the text, the sign of the factor is not identified, as one can see by observing that $-\gamma_i * (-F_t) = \gamma_i * F_t$. We handle this ambiguity by fixing the factor's coefficient of any particular time series to be positive so that for any given time period t, we are fixing the sign of the factor as well. Also, the scales of the factor loadings (the γ_is) and of the factor itself are not separately identified, as can be seen by noting that $\gamma_i/\eta * F_t * \eta = \gamma_i * F_t$ for any η. This problem is solved by normalizing the standard deviation of the innovations in the factor equation (3) to 1.

After solving these two identification issues, the first step is to set prior distributions for the parameters:

$$\gamma_i \mapsto N(\underline{\gamma},\underline{\Sigma}_\gamma);$$
$$(\rho_1,\rho_2) \mapsto N(\underline{\rho},\underline{\Sigma}_\rho)I_\rho(S);$$
$$(\varphi_{i,1},\varphi_{i,2}) \mapsto N(\underline{\varphi},\underline{\Sigma}_\varphi)I_\varphi(S);\text{ and}$$
$$\sigma_i^2 \mapsto IG(\underline{\alpha},\underline{\beta}).$$

In the previous expressions we use the symbol $\mapsto$ to denote "distributed as." A normal distribution for the intercepts and an inverse gamma (IG) for the variances are typical choices in Bayesian econometric models. Also note that for the φs and the ρs we impose a stationarity restriction, represented by an indicator variable that takes the value 1 if the parameter is inside the stationarity region S and 0 otherwise.

Starting with a guess for the parameter vector $\theta = \rho_1,\rho_2,(\gamma_1,\ldots,\gamma_n),(\varphi_{1,1},\ldots,\varphi_{1,n},\varphi_{2,1},\ldots,\varphi_{2,n})$, $(\sigma_1,\ldots,\sigma_n)$, from the following distributions, we sequentially

(a) sample the unobserved factors, $F_t \mid \{Y\}_t$, $\theta \mapsto N(F_t^*,P_t^*)$;[1]

(b) sample the ρs, $(\rho_1,\rho_2) \mid \{Y\}_t, \{F\}_t \mapsto N(\overline{\rho},\overline{\Sigma}_\rho)$; and

(c) for $i = 1,\ldots,n$, sample $\gamma_i \mid \{Y\}_t, \{F\}_t, \sigma_i, \varphi_{i,2} \mapsto N(\overline{\gamma}_i,\overline{\Sigma}_{\gamma,i})$, $\varphi_{i,1}, \varphi_{i,2} \mid \{Y\}_t, \{F\}_t, \sigma_i^2, \gamma_i \mapsto N(\overline{\varphi}_i,\overline{\Sigma}_{\varphi,i})$ and $\sigma_i^2 \mid \{Y\}_t, \{F\}_t, \gamma_i, \varphi_{i,1}, \varphi_{i,2} \mapsto IG(\overline{\alpha}_i,\overline{\beta}_i)$.

This sequential sampling is repeated several thousand times. At each step we condition on the previously sampled values for the parameters and the unobserved factor. We eliminated the first 1,500 draws to avoid having an influence from the initial conditions.

1. This step is the most involved. One must first apply the Kalman filter to the system in order to compute the mean and the covariance matrix for the unobserved factor at each point in time (see Kim and Nelson 1999, chap. 8, or Carter and Kohn 1994).

REFERENCES

Carter, C.K., and R. Kohn. 1994. On Gibbs sampling for state space models. *Biometrika* 81, no. 3:541–53.

Casella, George, and Edward I. George. 1992. Explaining the Gibbs sampler. *American Statistician* 46, no. 3:167–74.

Crone, Theodore M., and Alan Clayton-Matthews. 2005. Consistent economic indexes for the 50 states. *Review of Economics and Statistics* 87, no. 4: 593–603.

Kim, Chang-Jin, and Charles R. Nelson. 1999. *State-space models with regime switching: Classical and Gibbs-sampling approaches with applications.* Cambridge, Mass.: MIT Press.

Otrok, Christopher, Pedro Silos, and Charles H. Whiteman. 2003. Bayesian dynamic factor models for large data sets: Measuring and forecasting macroeconomic data. University of Iowa unpublished manuscript.

Otrok, Christopher, and Charles H. Whiteman. 1998. Bayesian leading indicators: Measuring and predicting economic conditions in Iowa. *International Economic Review* 39, no. 4:997–1014.

Stock, James, and Mark Watson., 1989. The revised NBER indexes of coincident and leading economic indicators. In *NBER macroeconomics annual 1989*, edited by Olivier J. Blanchard and Stanley Fischer. Cambridge, Mass.: MIT Press.

Tanner, Martin A., and Wing Hung Wong. 1987. The calculation of posterior distribution by data augmentation. *Journal of the American Statistical Association* 82, no. 398:528–40.