How Good Is What You've Got? DGSE-VAR as a Toolkit for Evaluating DSGE Models

MARCO DEL NEGRO AND FRANK SCHORFHEIDE
Del Negro is a research economist and assistant policy adviser in the Atlanta Fed's research department. Schorfheide is an associate professor at the University of Pennsylvania. *The authors thank Tom Cunningham, Pedro Silos, and Ellis Tallman for helpful comments.*

Dynamic stochastic general equilibrium (DSGE) models are becoming increasingly popular in central banking circles. The number of central bank–sponsored conferences on DSGE modeling and the amount of staff resources devoted to DSGE model development and estimation have risen dramatically over the past five years. This trend has affected monetary policy authorities around the globe, including the Federal Reserve System, the Bank of Canada, the European Central Bank, the Sveriges Riksbank, and the Reserve Bank of New Zealand. While few central banks are currently using DSGE models to generate forecasts and policy scenarios that provide the basis for interest rate decisions, many are contemplating doing so in the near future.

Part of the recent popularity of DSGE models is due to work by Smets and Wouters (2003), who document that a modified version of a New Keynesian model developed by Christiano, Eichenbaum, and Evans (2005) is able to track and forecast euro area time series as well as, if not better than, a vector autoregression (VAR) estimated with Bayesian techniques. While the empirical finding needs to be qualified, the results have had a considerable impact on how policymakers view DSGE models and have triggered efforts at many central banks to develop their own estimated DSGE model.[1]

In the 1990s the prevailing view among some policymakers was that DSGE models provide "good theory" to sharpen the understanding of business cycle fluctuations and to address fundamental policy questions: How is the stabilization of output and inflation through monetary policy actions related to the maximization of aggregate welfare? Should the central bank react to asset market fluctuations? How should monetary policy be conducted if the nominal interest rate is close to its zero lower bound? Should central banks of small open economies respond to exchange rate movements? Despite these benefits, many policymakers were skeptical that DSGE models could be used for quantitative data analysis, especially short- and medium-term forecasting and the projection of macroeconomic aggregates under alternative interest rate scenarios.

At the same time most academic macroeconomists were still reluctant to use the kind of econometric techniques that enable careful documentation of the time series fit of a dynamic model. Many DSGE models impose very strong restrictions on actual time series and are rejected against less restrictive specifications such as VARs. Even though it has long been known that DSGE models can be estimated, their apparent misspecification was used as an argument in favor of informal calibration approaches, along the lines of Kydland and Prescott (1982).

In recent years econometricians have developed frameworks that formalize certain aspects of the calibration approach by taking the possibility of model misspecification explicitly into account without abandoning the tradition of probabilistic modeling initiated by Haavelmo (1944). In particular, several authors, including DeJong, Ingram, and Whiteman (2000), Schorfheide (2000), Otrok (2001), and Fernández-Villaverde and Rubio-Ramírez (2004), documented that Bayesian methods can be used in an insightful manner to estimate and evaluate DSGE models.

In the 1990s many policymakers were skeptical that DSGE models could be used for quantitative data analysis, especially short- and medium-term forecasting and projecting macroeconomic aggregates under alternative interest rate scenarios.

Smets and Wouters (2003) applied the newly developed Bayesian methods to a DSGE model with enough nominal and real frictions that their specification had a good chance of fitting major aggregate time series in a traditional macroeconometric sense. In fact, the model development process was a productive synthesis of academic-style DSGE modeling and econometric model building. Technology and monetary policy shocks—the most common driving processes in theoretical models—were augmented by a list of shocks that to a large extent were chosen to pick up serial correlation of the wedges in intra- and intertemporal equilibrium conditions, the DSGE-model equivalent of regression residuals.

It is no surprise that central banks have paid a great deal of attention to Smets and Wouters's results and now devote significant resources to developing and estimating their own DSGE models. The best of two worlds appears within reach: a model that is well founded from a theoretical perspective and at the same time in tune with the empirical evidence so that it can deliver reliable forecasts and a coherent interpretation of past and current economic events.

Now that policy institutions are beginning to take the quantitative implications of DSGE models seriously, there is a need for robust evaluation procedures. The head of any central bank's research department that has just built a DSGE model for policy analysis and forecasting would want to know: How good is this model? Is it reliable enough so that it can be used for policy advice? Does the model need to be improved, for instance, by explicitly modeling labor market frictions, credit market imperfections, or information asymmetries? In principle, time will tell. As the model is used on a regular basis, analysts will discover ex post its strengths and weaknesses and point to directions for improvement. However, this real-time learning process is potentially slow and costly. Hence, it is important to subject the model to evaluation procedures that can signal deficiencies ex ante—that is, on the basis of the information currently available.

In an attempt to make the fit and forecasting performance of DSGE models comparable to a VAR, the structural models have been augmented by features that appear ad hoc and lack micro foundations. For instance, price stickiness is often introduced by assuming that only a fraction of firms are able to reoptimize their nominal prices. By itself this mechanism might not generate enough persistence to

explain the high autocorrelation of inflation rates in the data. Hence, researchers often add the assumption that those firms that do not reoptimize their prices can costlessly adjust their old prices by last period's inflation rate. Alternatively, the model could be augmented by serially correlated price markup shocks, which might reflect either time variation in the substitution elasticities between differentiated goods or the market structure of an industry. However, there is a trade-off for incorporating ad hoc propagation mechanisms or exogenous shocks into the model. On the one hand, model fit with respect to historical data is typically improved. On the other hand, it is questionable whether the ad hoc modifications are invariant to policy experiments. Although in the absence of large historical variation in monetary policy the invariance property is to some extent difficult to assess, the trade-off should serve as a word of caution and steer modelers toward parsimony and internal propagation mechanisms that are supported by microeconometric evidence.

In sum, there is a need for DSGE model evaluation procedures. This article reviews an evaluation procedure recently proposed in Del Negro and Schorfheide (2004) and Del Negro et al. (2004). The article first describes the DSGE model and the data used in the empirical application. The article next shows how the linear DSGE model can be nested in a VAR and reviews a procedure that is able to systematically relax the cross-coefficient restrictions imposed on the VAR by the DSGE model. The resulting DSGE-VAR specification is used as a tool to evaluate a version of the Smets and Wouters (2003) model. The analysis considers to what extent the DSGE model restrictions must be relaxed in order to optimize the fit of the DSGE-VAR and then uses the framework for comparisons of different DSGE model specifications. The article then describes some in- and out-of-sample results obtained with this procedure.[2]

The DSGE Model

The DSGE model used in the forecasting exercise is described in detail in Del Negro et al. (2004). The model is a slightly modified version of the DSGE model in Smets and Wouters (2003), which is in turn based on work of Christiano, Eichenbaum, and Evans (2005). Here we provide a brief and nontechnical overview of the model.

The model contains several nominal and real frictions. Nominal price and wage stickiness is modeled as in Calvo (1983). Firms (households) are monopolistic suppliers of a differentiated good (labor). In any period there is a chance that any given firm (household) may not be able to reset prices (wages). The prices (wages) of these firms (households) grow proportionally to the previous period's inflation. (This proportional growth is referred to as indexation in the remainder of the article.)

On the real side, the model features endogenous capital accumulation, adjustment costs to investment, and variable capital utilization. Households' preferences display habit persistence in consumption, and the utility function is separable in terms of consumption, leisure, and real money holdings. Fiscal policy amounts simply to balancing the budget in all periods. Monetary policy follows an interest feedback rule, in which the target federal funds rate depends on the rate of inflation and on the discrepancy between actual and trend output and adjustment to the target is gradual.

---

1. The empirical findings need to be qualified because Smets and Wouters worked with detrended data and thus did not use the most favorable prior for the Bayesian VAR and because VARs are not universally favored as a forecasting benchmark.
2. The results shown in this article are variations or extensions of those discussed in Del Negro et al. (2004).

As in Smets and Wouters (2003), the model economy is subject to a large number of shocks: technology, discount rate, leisure preference, price markup, investment efficiency, monetary policy, and government spending. Technology shocks are assumed to be permanent and common to all firms. Discount rate and leisure preference shocks shift households' utility; the first affects the household's willingness to substitute over time, and the latter the household's willingness to supply labor. So-called price markup shocks change the degree of substitutability among differentiated goods and in turn affect markups and the rate of inflation. Investment efficiency shocks alter the rate of transformation between consumption and investment goods and serve as proxies for changes in the relative price of investment goods. Finally, both monetary policy and government spending shocks have a standard interpretation. All shocks are assumed to follow an autoregressive process of order one (in the case of technology, this assumption applies to the growth rate of technology) with the exception of monetary policy shocks, which are independently distributed over time.

In recent years econometricians have developed frameworks that formalize certain aspects of the calibration approach by taking the possibility of model misspecification explicitly into account without abandoning the tradition of probabilistic modeling.

The Data and the VAR Setup

The empirical analysis is based on quarterly U.S. observations that include both real and nominal series. The real variables are per capita real output, investment, consumption, hours per capita, and wages. The nominal variables are inflation and the interest rate. All data are obtained from Haver Analytics; Haver's abbreviations are in italics. Consistent with much of the real business cycle literature, this analysis treats consumption of durable goods (*CD*) as investment rather than consumption. Therefore, investment is defined as gross private domestic investment plus consumption of durables. Per capita real output, investment, and consumption are obtained by dividing the nominal series (*GDP*, *C* – *CD*, and *I* + *CD*, respectively) by the population sixteen years and older (*LN16N*) and deflating using the chained-price GDP deflator (*JGDP*). The real wage is computed by dividing compensation of employees (*YCOMP*) by total hours worked and the GDP deflator. Note that compensation per hour, which includes wages as well as employer contributions, accounts for both wage and salary workers and proprietors. The measure of hours worked is computed by taking total hours worked reported in the national income and product accounts (NIPA) (annual frequency) and interpolating it using growth rates computed from hours of all persons in the nonfarm business sector (*LXNFH*). Hours worked are divided by population to convert them into per capita terms. The analysis therefore uses a broad measure of hours worked that is consistent with its definition of both wages and output in the economy. Inflation rates are defined as log differences of the GDP deflator (*JGDP*) and converted into annualized percentages. The nominal rate corresponds to the effective federal funds rate (*FFED*), averaged within each quarter, also in percent.

As mentioned in the introduction, the DSGE model is nested in a more flexible vector autoregressive specification. The DSGE model features a stochastic trend, driven by the permanent technology shock. Real per capita output, consumption, investment, and the real wage are nonstationary and grow at the same rate in the long run. These nonstationary variables enter the VAR in growth rates, while the variables that are stationary according to the DSGE model—namely, per capita hours, inflation, and the nominal interest rate—enter the VAR in levels. All growth rates are computed

using quarter-to-quarter log differences and then multiplied by 100 to convert them into percentages. To take into account the fact that the nonstationary variables all move together in the long run according to the DSGE model, error-correction terms are also introduced into the VAR so that effectively we are estimating a vector error correction model (VECM). Importantly, to maintain the consistency between the VAR and the DSGE model, the coefficients in the cointegrating relationships are constrained to be those implied by the DSGE model—that is, real per capita output, consumption, investment, and the real wage are assumed to grow at the same rate in the long run in the VAR as well. The error correction terms are therefore given by consumption minus output, investment minus output, and the real wage minus output, respectively, all in logarithms.

The analysis uses observations from 1954Q4 to 2004Q1. The first four observations are used to initialize the lags of the VAR. The recursive estimation results and the pseudo-out-of-sample forecasts are based on a rolling sample of 120 observations (thirty years) starting in 1956Q1. Specifically, the rolling sample works as follows. We estimate the model on the sample 1956Q1–1985Q4, produce forecasts, and then shift the sample one quarter ahead and repeat the exercise. Therefore we have arguably a sample large enough to estimate the model as well as enough forecasts (fifty-eight) to assess the accuracy of out-of-sample predictions.

DSGE-VAR: A Brief Description of the Procedure

The short and informal description of the DSGE-VAR procedure in this section is intended to motivate the DSGE-VAR specification and to provide some intuition on how it can be used to estimate and evaluate DSGE models.[3] It has long been recognized (for example, Sims 1980) that a tight relationship exists between dynamic equilibrium models and VARs. Imagine the following thought experiment, where for the moment the vector of DSGE model parameters is fixed. We generate 1 million observations from the DSGE model—that is, we generate a sequence of shocks (monetary policy, technology, etc.), feed them trough the DSGE model, and obtain artificial data. Next, we estimate a VAR with p lags on these artificial data. If the DSGE model is covariance stationary, then the estimated VAR provides an approximation to the DSGE model with the property that its first p autocovariances are equivalent to the first p autocovariances of the DSGE model. By including more and more lags we can in principle match more and more autocovariances and increase the accuracy of the VAR approximation of the DSGE model. Now imagine that the data generation is repeated using different parameter values for the DSGE model. As long as the DSGE model parameter space is small compared to the VAR parameter space, a restriction function can be traced that maps the DSGE parameters into a VAR parameter subspace. Hence, estimating a DSGE model is (almost) like estimating a VAR with cross-equation restrictions.

Instead of dogmatically imposing the cross-coefficient restrictions implied by the DSGE model on the VAR, we will allow for deviations. The overall magnitude of these deviations is controlled by a hyperparameter, λ. Roughly speaking, if $\lambda = \infty$, then the restrictions are strictly enforced, whereas if $\lambda = 0$, the restrictions are completely

3. The section—and the whole article for that matter—purposely does not contain a formal treatment of the procedure. The latter is provided in Del Negro and Schorfheide (2004) and Del Negro et al. (2004). The appendix in Del Negro and Schorfheide (2004) discusses computational details for readers who are interested in implementing the procedure. Gauss and Matlab versions of the codes are available at www.econ.upenn.edu/~schorf/research.htm.

ignored in the estimation of the VAR parameters. To implement this idea formally, we use a Bayesian approach. In general terms, Bayesian methods are a collection of inference procedures that combine initial information about parameters with sample information in a logically coherent manner by use of Bayes's theorem. Both prior and postdata information are represented by probability distributions. In this particular application, the prior consists of a continuous probability distribution for the VAR coefficients that is centered at the DSGE model implied restrictions.

The best of two worlds appears within reach: a model that is well founded theoretically and can deliver reliable forecasts and a coherent interpretation of past and current economic events.

The hyperparameter λ scales the covariance matrix of the prior: If λ is large the variance is small, and most of the prior mass on the VAR coefficients concentrates near the DSGE model restrictions. Vice versa, if λ is small the prior on the VAR coefficients is diffuse. The prior is combined with the likelihood function to form the posterior distribution, which summarizes the postdata information about the VAR parameters. The larger λ is, the more the posterior shifts toward the DSGE model restrictions and the less the restrictions are relaxed in the estimation. We refer to the resulting vector autoregressive specification as DSGE-VAR. In the application the DSGE model depends on unknown parameters as well. It turns out that these parameters can be jointly estimated together with the VAR parameters by, loosely speaking, projecting the VAR coefficient estimates back onto the DSGE model restrictions.

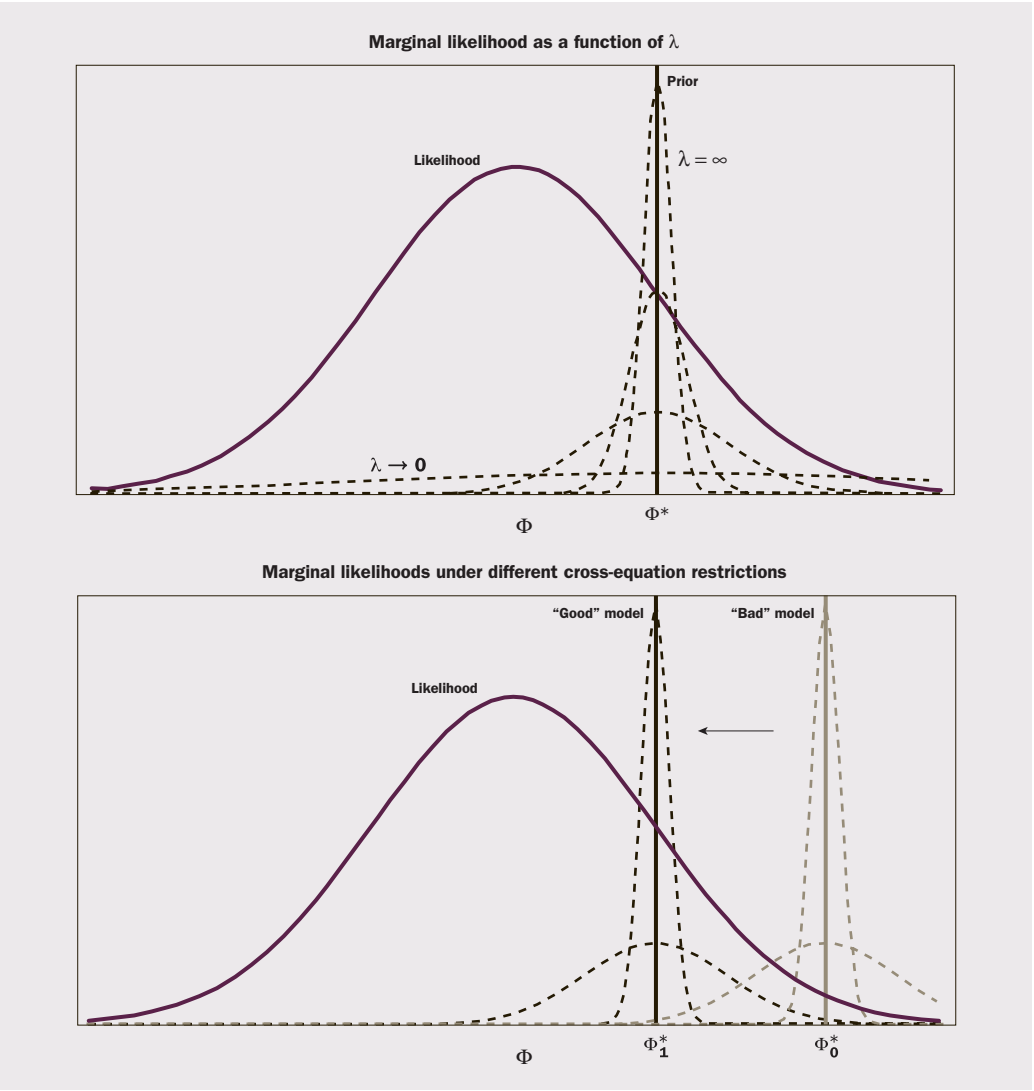
Both fit and forecasting performance suffer whenever the DSGE prior is either too tight or too loose. The fact that fit improves as the cross-equation restrictions are relaxed—that is, as λ decreases from infinity—indicates that these restrictions are at odds with the data in some dimensions. In the procedure proposed in Del Negro et al. (2004), an estimate of λ is used as a way to evaluate DSGE models. That is, the evaluation procedure hinges on the following question: How much must the cross-equation restrictions be relaxed to obtain the best-fitting model? The next section elaborates on why the answer to this question can shed light on some of the issues described in the introduction.

Why Does λ Tell Us How Good a DSGE Model Is?

In this section, a simple chart provides some intuition for the DSGE-VAR procedure. The first panel of Figure 1 plots the likelihood of the VAR as a function of the VAR parameter Φ. For the sake of exposition the multidimensional VAR parameter space is collapsed onto the real line. Assume that the DSGE model restrictions imply that the VAR parameter equals Φ^*. The remaining lines represent the DSGE prior for different values of λ. All these priors are centered at the cross-equation restrictions Φ^*. For $\lambda = \infty$ the prior puts all its mass on Φ^*. As λ decreases, the prior mass is spread out further away from the cross-equation restrictions. For λ approaching zero the prior becomes nearly flat. In a Bayesian setting a model consists of a likelihood function and a prior distribution. By varying the hyperparameter λ from infinity to zero we are essentially creating a continuum of models with the VAR approximation of the DSGE model at one end and an unrestricted VAR at the other end.

We adopt a measure of model fit that has two dimensions: goodness of in-sample fit on the one hand and a penalty for model complexity or degrees of freedom on the other hand. In a Bayesian framework such a measure is provided by the so-called marginal data density, which arises naturally in the computation of posterior model odds. The marginal data density is simply the integral of the likelihood taken according to the prior

Figure 1
Marginal Likelihoods and DSGE Priors



distribution—that is, the weighted average of likelihood where the weights are given by the prior. We then ask the following question: How does this measure of fit change as λ decreases from infinity to zero? We refer to the mapping from λ to the marginal data density as the posterior distribution of λ. Indeed, if we view λ as a hyperparameter and put a flat prior on it, this mapping characterizes a posterior distribution of λ.

Suppose one writes an oversimplified DSGE model, whose cross-equation restrictions are grossly at odds with the data. The first panel of Figure 1 clearly shows that if Φ^* is far in the tails of the likelihood, any prior that is very tight around Φ^* will have low marginal likelihood. As λ is decreased, the weight on parameters in the calculation of the data density that are associated with a high likelihood increases. Hence, small values of λ have large posterior weights. Notice, however, as λ approaches zero, the computation of the data density involves more parameter values for which

the likelihood function is essentially zero. Hence, one expects the posterior density of λ to fall eventually.

Now imagine improving the model by adding a number of frictions that generate more realistic cross-equation restrictions. One expects that the posterior distribution of λ will concentrate more mass on large values of the hyperparameter λ. The reasoning is as follows. Having better cross-equation restrictions means that Φ^* moves closer to the likelihood peak, as shown in the second panel of Figure 1. As a consequence, relatively tight priors will deliver a higher marginal likelihood than loose priors. As the posterior distribution of λ shifts to the right, its mode—the value $\hat{\lambda}$ that maximizes the marginal likelihood—will increase. The remainder of the article provides concrete examples of how the posterior distribution of λ can shift as the underlying DSGE model changes.

The fact that the DSGE model's fit improves as the cross-equation restrictions are relaxed indicates that these restrictions are at odds with the data in some dimensions.

What is the appeal of this procedure relative to the current practice in the literature? Following the work of Smets and Wouters (2003), a standard approach for evaluating the overall fit of a DSGE model is to compare its marginal data density (see definition above) with that of a Bayesian VAR (BVAR). Although most VARs used in practice are not equipped with a DSGE model prior—most researchers use either a version of the Minnesota prior (see Doan, Litterman, and Sims 1984) or a training sample prior—the problems arising in such a comparison can be discussed in the context of our framework. Current practice is to consider two extremes: On the one hand, $\lambda = \infty$ represents the DSGE model, and on the other hand, a small value $\lambda = \underline{\lambda}$ is a proxy for the BVAR that serves as a benchmark in the evaluation exercise. By using a very diffuse prior on one or more of the BVAR parameters—that is, choosing a low $\underline{\lambda}$—one can make the marginal likelihood of the BVAR arbitrarily small. So one can always make the BVAR lose the horse race with the DSGE model by choosing, often unconsciously, a diffuse prior. At the same time, the VAR coefficient estimates simply converge to the maximum likelihood estimates as $\underline{\lambda}$ approaches zero.

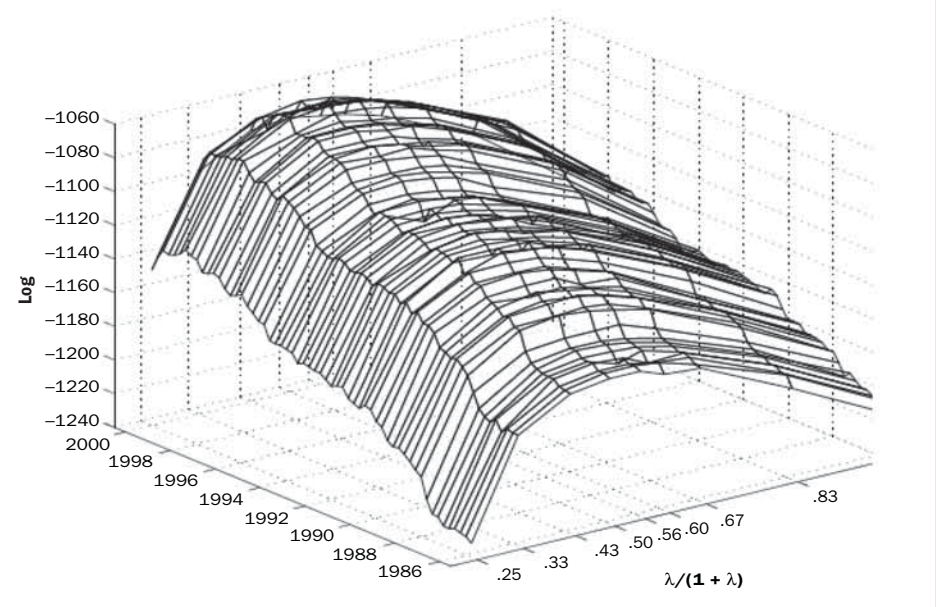
The sensitivity of posterior odds comparisons to seemingly innocuous changes in the prior distribution of the benchmark model implies that posterior odds of DSGE versus VAR models are not a robust way to address the question, How good is my DSGE model? Our DSGE-VAR framework imposes some rigor on the construction of the prior for the VAR; we emphasize that it is important to look at marginal data densities for an entire range of λ values instead of just two endpoints, one of which is typically chosen in a fairly arbitrary manner. As soon as we allow for intermediate values of λ, minor changes in model specifications are less likely to affect the answer to the question, Is there evidence of misspecification? Indeed, in Del Negro et al. (2004) and in the present article, we show that the overall shape of the posterior distribution of λ is quite a robust feature of the DSGE-VAR procedure.

A Look at the Data: The Posterior Distribution of λ over Time

The posterior distribution of λ is one of the main objects of interest in our empirical analysis: For any given sample it provides information on the degree of misspecification of the DSGE model. Since we estimate the DSGE-VAR in our empirical analysis not just once but essentially fifty-eight times based on the rolling samples, we can study how the posterior distribution of λ evolves over time.

Figure 2 shows the evolution of the posterior distribution of λ over time using a three-dimensional plot. For each of the fifty-eight rolling windows—the first ending

Figure 2
Marginal Likelihood as a Function of λ over Time



in 1985Q4 and the last one ending in 2000Q1—the marginal likelihood is computed for the following values of λ in the [0.33, ∞] interval (where 0.33 is the smallest value of λ that generates a proper prior for the VAR parameters): 0.33, 0.5, 0.75, 1, 1.25, 1.5, 2, 5, ∞. The *x* axis of Figure 2 shows the values of λ, which for expositional purposes are rescaled to be in the [0, 1] interval—that is, the value of λ/(1 + λ). The *y* axis shows the ending period of the rolling window, and the *z* axis shows the corresponding value of the logarithm of the marginal likelihood. Therefore, for any given rolling window ending between 1985Q4 and 2000Q1, the plot shows how the marginal likelihood of DSGE-VAR(λ) evolves as a function of λ.

The shape of the three-dimensional plot in Figure 2 is consistent with what we would expect. For any given window, the marginal likelihood initially increases with λ. Recall from Figure 1 that if the cross-equation restrictions are not too far in the tail of the likelihood—that is, if the DSGE model misspecification is not too large—tightening the DSGE prior leads to an improvement in the marginal likelihood. However, as the DSGE prior concentrates around the cross-equation restrictions, the marginal likelihood starts to decrease. We interpret this as evidence of misspecification because it suggests that relaxing the cross-equation restrictions improves the DSGE-VAR's fit.

The fact that the three-dimensional plot in Figure 2 looks like a tunnel indicates that the shape of the posterior distribution of λ is very robust over the sample period. Interestingly, this result is in contrast with other approaches to assessing the DSGE model's fit, as we will presently show. The tunnel is upward sloping in the time (*y*) dimension: For any given value of λ, the marginal likelihood tends to increase as the rolling window shifts forward. This phenomenon possibly reflects what has been dubbed the Great Moderation (Stock and Watson 2002): After the mid-eighties the volatility of

key macroeconomic variables has dropped sharply. Consequently, the predictability of these variables has increased, leading to an increase in the marginal likelihood. The ranking of fit as a function of λ is fairly stable over time with values of λ in the neighborhood of 1 (that is, $\lambda/(1+\lambda)$ around 0.5) always outperforming the two extremes of the interval—namely, DSGE-VARs with either a very loose or a very tight prior.

The relative ranking of the extremes of the λ interval—namely, the "loose prior" ($\lambda = 0.33$) versus the "degenerate prior" ($\lambda = \infty$) model—is not very robust, however. Comparing the fit of the DSGE model with that of a VAR with a loose prior—an approach that is often used in the literature—leads to conclusions that change dramatically over the sample period, even though the overall shape of the posterior distribution of λ is roughly the same. Since this pattern is difficult to assess from Figure 2, the two charts in Figure 3 show slices of the tunnel, one taken at the beginning (1985Q4) and one at the end (2000Q1) of the rolling sample. The two charts in Figure 3 also contain a comparison across different models, which is the subject of the next section. For now, we focus on the Baseline model (the heavier black line), which plots the same numbers that are in Figure 2. The figure shows that at the beginning of the rolling sample, DSGE-VAR(∞) outperforms the VAR with a loose prior; the ranking is reversed at the end of the sample. The log difference between the marginal likelihood of DSGE-VAR(∞) versus DSGE-VAR(0.33) is 19 at the beginning of the rolling sample and –4 at the end. Taken literally, these differences imply posterior odds that are in one case decisively in favor of, and in the other case against, the DSGE model. Once again, the overall shape of the posterior distribution is roughly the same in both charts. In fact, in both cases the two extremes are in tails of the posterior distribution of λ, their posterior odds relative to the best-fitting DSGE-VAR being negligible.

As the features of the DSGE model change, so do the cross-equation restrictions that the model imposes on its VAR representation.

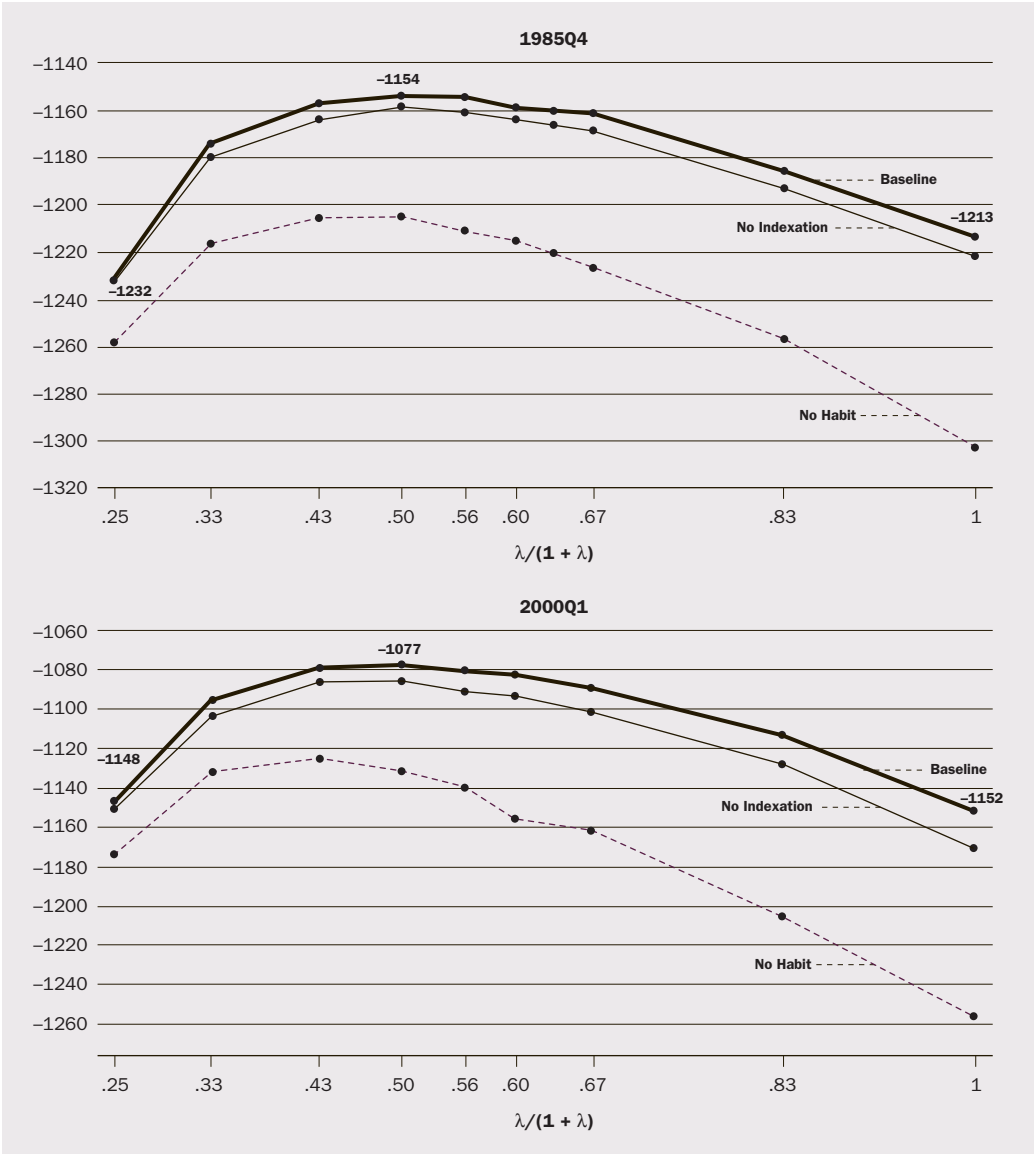
This last observation implies that the VAR with a loose prior may not be the right reference model to use for impulse response comparison because its fit is sometimes worse relative to that of the model being analyzed (the DSGE model). Certainly, both Figures 2 and 3 show that the fit of the VAR with a loose prior is always much worse than that of the best-fitting model DSGE-VAR($\hat{\lambda}$). This result suggests that the latter provides a more reliable benchmark.

Model Comparisons

In this section we use the DSGE-VAR procedure to compare across DSGE models. As the features of the DSGE model change, so do the cross-equation restrictions that the model imposes on its VAR representation. For the reasons discussed in the article's introduction, we are interested in determining which model features are truly important and which are not.

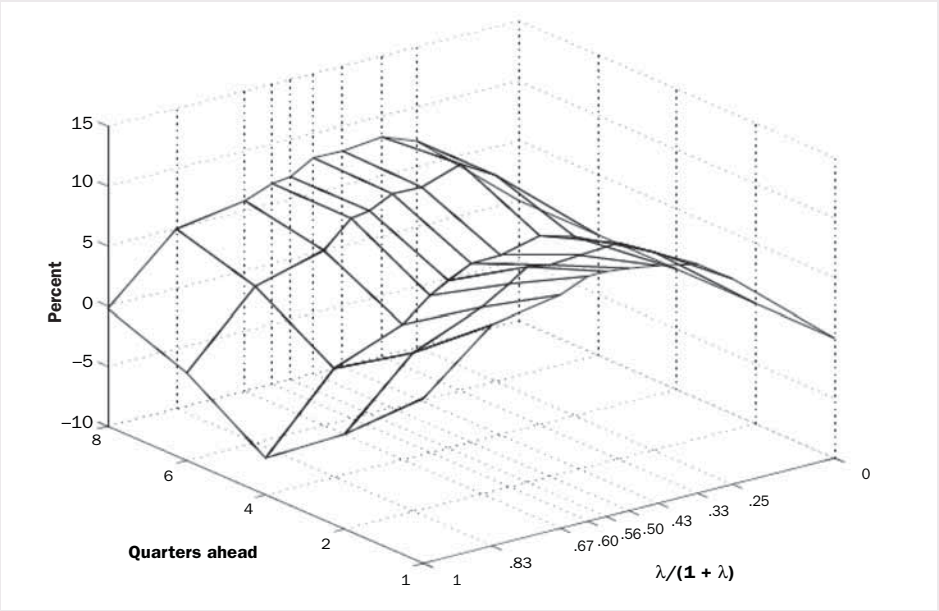
In Del Negro et al. (2004) we consider two alternatives, also shown in Figure 3, to the so-called Baseline model discussed so far. The No Indexation specification (the thinner black line) eliminates price and wage indexation to last period's inflation. Price and wage indexation is often viewed as being somewhat at odds with microeconomic evidence and therefore considered not truly structural. The other specification, called No Habit (the dashed purple line), eliminates habit persistence in preferences. Some of the literature has argued that these features are needed in order to fit the data. Here we use the DSGE-VAR procedure to assess whether this is the case. We also want to learn whether the conclusions from the procedure are robust across samples.

Figure 3
Marginal Likelihood as a Function of λ for Different DSGE Models



We are interested in studying how the shape of the posterior distribution of λ changes across models. We argued previously that the posterior distribution of λ shifts to the left if the misspecification of the cross-equation restrictions (which we referred to as Φ^*) increases. Figure 3 shows that this shift indeed occurs for the No Habit model, regardless of the sample. Relative to the Baseline model, the marginal likelihood is much lower for any value of λ. Most importantly, the posterior mass clearly shifts to the left, toward lower values of λ (looser prior). This is not so much the case, however, for the No Indexation model. The marginal likelihood is slightly lower for the No Indexation relative to the Baseline model for any λ, indicating that the fit worsens, but the posterior distribution does not show any appreciable shift to the left.

Figure 4
Forecasting Accuracy as a Function of λ: Baseline Model



We interpret these findings as strong evidence that the habit persistence in preferences substantially improves the fit of the DSGE model. Therefore, those who believe that habit persistence is not a structural feature may have to introduce an alternative mechanism that delivers similar effects: Simply eliminating habit persistence comes at a significant cost in terms of fit. On the contrary, the evidence in favor of price and wage indexation is not nearly as strong.

Forecasting Results

This section presents some forecasting results obtained with the models discussed above. In particular, we want to find out to what extent the in-sample results shown so far carry over to the pseudo-out-of-sample comparison. Figure 4 shows a multivariate forecasting statistic for DSGE-VAR(λ) relative to the unrestricted VAR for forecasting horizons one through eight quarters ahead. The multivariate forecasting statistic is a summary measure of forecasting accuracy. Loosely speaking, this multivariate measure can be seen as a weighted average of the root mean square error for the individual variables. Differently from a simple weighted average, however, this measure also takes into account the correlation in the forecast errors. The *z* axis in Figure 4 reports the percentage gain in multivariate forecasting accuracy relative to the unrestricted VAR. As in Figure 2, we rescale λ to be in the [0, 1] interval. Thus, the *x* axis shows the value of $\lambda/(1+\lambda)$, and the *y* axis shows the forecast horizon.

Just as in Figure 2, the three-dimensional plot in Figure 4 is also tunnel-shaped. In other words, forecasting accuracy is an inverted U-shaped function of λ for all forecast horizons. Consistently with the in-sample results, forecasting performance is maximized whenever the DSGE prior is neither too loose nor too tight. The magni-

Table
Out-of-Sample Root Mean Square Errors: Percentage Improvement Relative to VAR

		Forecast horizon (quarters)				
		1	2	4	6	8
Y	DSGE-VAR($\hat{\lambda}$)	16.3	14.1	12.5	13.5	13.6
	DSGE-VAR(∞)	0.9	–17.6	–56.5	–82.5	–102.9
	VAR, RMSE	*0.67*	*0.97*	*1.68*	*2.38*	*2.98*
C	DSGE-VAR($\hat{\lambda}$)	–6.8	–7.6	7.1	16.6	21.5
	DSGE-VAR(∞)	–15.7	–21.4	–0.8	11.3	12.0
	VAR, RMSE	*0.42*	*0.62*	*1.06*	*1.56*	*2.03*
I	DSGE-VAR($\hat{\lambda}$)	17.8	8.0	–5.0	–11.5	–17.2
	DSGE-VAR(∞)	–4.2	–41.2	–101.0	–135.3	–157.8
	VAR, RMSE	*2.67*	*3.98*	*6.59*	*9.14*	*11.45*
H	DSGE-VAR($\hat{\lambda}$)	10.0	10.9	–0.6	–0.0	0.7
	DSGE-VAR(∞)	–13.6	–37.9	–95.4	–116.5	–127.2
	VAR, RMSE	*0.58*	*0.92*	*1.56*	*2.26*	*2.88*
W	DSGE-VAR($\hat{\lambda}$)	8.2	11.7	11.1	14.9	18.4
	DSGE-VAR(∞)	6.7	12.7	18.1	27.0	36.6
	VAR, RMSE	*0.65*	*1.06*	*1.72*	*2.28*	*2.82*
Inflation	DSGE-VAR($\hat{\lambda}$)	10.7	10.9	22.9	31.0	36.6
	DSGE-VAR(∞)	8.4	4.2	10.4	21.1	29.6
	VAR, RMSE	*0.25*	*0.47*	*0.98*	*1.68*	*2.42*
R	DSGE-VAR($\hat{\lambda}$)	27.3	23.4	9.2	7.0	9.1
	DSGE-VAR(∞)	27.7	17.8	3.2	8.2	17.1
	VAR, RMSE	*0.68*	*1.14*	*1.63*	*2.11*	*2.64*
Multivariate statistic	DSGE-VAR($\hat{\lambda}$)	11.0	8.8	6.1	9.4	9.4
	DSGE-VAR(∞)	3.8	–2.1	–6.9	–2.7	–0.2
	VAR, RMSE	*0.68*	*0.23*	*–0.18*	*–0.47*	*–0.65*

tude of the relative improvement in forecasting accuracy varies over the forecasting horizon. But for any given forecasting horizon, the values of λ that maximize forecasting performance are those in the neighborhood of 1 (that is, $\lambda/(1 + \lambda) = 0.5$), and roughly correspond to those that maximize in-sample fit.

The table takes a closer look at the forecasting performance of the unrestricted VAR, DSGE-VAR($\hat{\lambda}$), and DSGE-VAR(∞). For each of the seven variables and for the multivariate statistic, the table shows the root mean square error (RMSE) of the unrestricted VAR as well as the percentage improvement in forecasting accuracy (whenever positive) of DSGE-VAR($\hat{\lambda}$) and DSGE-VAR(∞) relative to the VAR. For DSGE-VAR($\hat{\lambda}$) the value of $\hat{\lambda}$ is chosen ex ante for each rolling sample on the basis of the marginal likelihoods shown in Figure 2. The precise value changes from sample to sample but is always in the neighborhood of 1, as one can see from Figure 2. For those variables that enter the estimation in growth rates (output, consumption, investment, and the real wage), as well as for inflation, we focus on cumulative forecasts. Therefore, for forecast horizons beyond one quarter, the forecast errors measure the cumulative error in forecasting inflation over, say, the next two years, as opposed to the error in forecasting the variable two years ahead. For instance, an eight-quarter-ahead error of 2 percent in

forecasting consumption implies that the model makes a mistake of 50 basis points (annualized) in forecasting average consumption growth in the next two years.

The table shows that for most variables and forecasting horizons the DSGE-VAR($\hat{\lambda}$) improves over the unrestricted VAR. This is certainly the case for the multivariate statistic, as already mentioned in the discussion of Figure 4. Short-run consumption forecasts and long-run investment forecasts are an exception. Interestingly, there seems to be a trade-off between forecasting consumption and investment. This trade-off reflects the fact that all three models considered in the table are error-correction models with the same long-run cointegrating restrictions on output, consumption, investment, and the real wage. Since these cointegrating restrictions are at odds with the data, accurate forecasts for some of these variables result in inaccurate forecasts for others given that not all series grow proportionally in the long run as the model predicts. Another manifestation of this phenomenon is the fact that DSGE-VAR(∞) outperforms the other two models in forecasting the real wage but performs very poorly in forecasting both output and investment, especially in the long run. In summary, the fact that the DSGE model imposes these long-run cointegrating restrictions results in a serious limitation of its forecasting ability. To the extent that DSGE-VAR inherits the same long-run restrictions, its accuracy suffers as well.

For the remaining variables, DSGE-VAR($\hat{\lambda}$) is roughly as accurate as the unrestricted VAR in terms of hours per capita, while DSGE-VAR(∞) is far worse, especially in the long run. Conversely, DSGE-VAR(∞) performs well in terms of the nominal variables, inflation and the interest rate. For inflation its forecasting accuracy is slightly inferior to that of DSGE-VAR($\hat{\lambda}$) and far superior to that of the unrestricted VAR. For the nominal interest rate, DSGE-VAR(∞) outperforms DSGE-VAR($\hat{\lambda}$) for longer forecast horizons, but in the short run the two models have roughly the same forecasting performance.

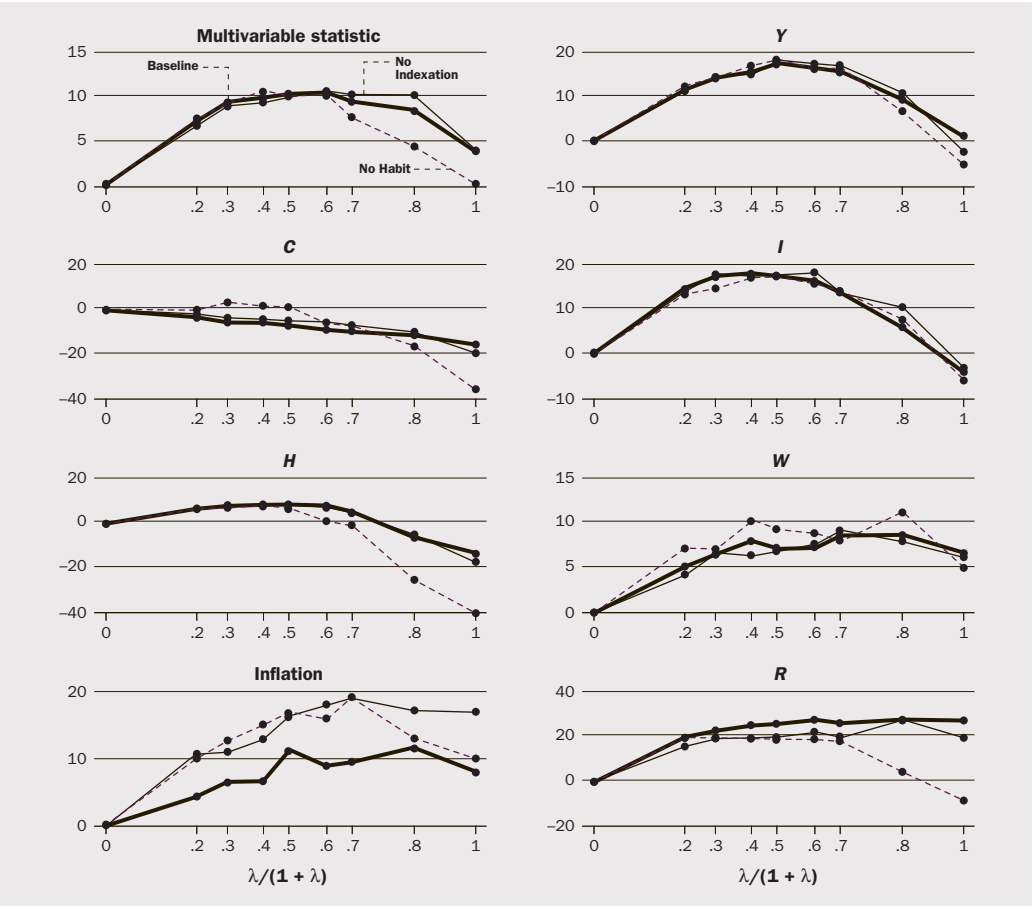
We conclude the section with a comparison of the out-of-sample forecasting performance across models. For each of the three models discussed so far (Baseline, No Indexation, No Habit), Figure 5 shows the one-quarter-ahead percentage improvement in RMSEs relative to the unrestricted VAR for all seven variables, as well as the improvement in the multivariate forecast statistic, as a function of λ. The focus on one-period-ahead forecasting accuracy facilitates the comparison with the results in Figure 3, which were based on the marginal likelihood.

The results in Figure 5 agree in a number of dimensions with those in Figure 3. The multivariate statistic plot, for instance, indicates that forecasting accuracy worsens considerably for the No Habit model as the DSGE prior becomes too tight. The plots for the individual variables show that for high values of λ the No Habit model performs worse than the other two models not only for consumption, as expected, but also for hours and the nominal interest rate. For other variables, however—notably, investment, real wages, and inflation—the No Habit model performs as well as the other two models.

Consistent with the overall message from the model comparison based on the marginal data densities, the No Indexation and Baseline models perform roughly as well in terms of the multivariate statistic. The forecasting performance of the two models is pretty much the same for most individual variables. One interesting exception is inflation, where the No Indexation model clearly forecasts better than the Baseline model. In summary, the out-of-sample exercise confirms the finding, consistent with our earlier discussion, that there is no strong evidence in favor of including wage and price indexation in the DSGE model.

In some other dimensions, however, the results in Figure 5 cast some doubt on the model comparison based on the marginal likelihood. For instance, the shape of the marginal likelihood curves as a function of λ for the No Indexation and Baseline models

Figure 5
Forecasting Accuracy as a Function of λ for Different DSGE Models



were very similar. Yet the marginal likelihood comparison, if taken literally, suggests that the No Indexation model should be at a loss relative to the Baseline model in terms of fit. This pattern does not emerge from the out-of-sample model comparison, however. Likewise, for low values of λ the difference in marginal likelihoods between the No Habit and the other two models is narrower than for large values of λ but still quite large. In the out-of-sample comparison, however, the three models seem to perform equally well in terms of the multivariate statistic for low values of λ.

DSGE-VAR as a Reference Model

In the previous sections, we argued that the posterior distribution of λ provides a robust measure of overall fit for the DSGE model. However, we often want to know more than just whether a DSGE model's fit is good or not. If the model fails—that is, if the λ posterior does not peak around a large value—we want to know in which dimensions it has to be improved. Although the forecast results provide us with information about the accuracy with which individual variables can be predicted by the model, they do not document how well the structural model captures comovements.

Comovements and the propagation of structural shocks can be illustrated with impulse response functions. While it is straightforward to compute impulse responses

from an estimated DSGE model, finding an appropriate benchmark to which these responses can be compared is more difficult. Many authors, including Nason and Cogley (1994), Rotemberg and Woodford (1997), Schorfheide (2000), and Christiano, Eichenbaum, and Evans (2005), have compared impulse responses from a DSGE model to responses obtained from a VAR. Such a comparison faces two challenges. First, for the VAR to be a meaningful benchmark, it has to fit the data better, in an econometric sense, than the DSGE model. Second, the VAR has to be expressed in terms of structural shocks—that is, technology shocks, monetary policy shocks, and so forth—rather than reduced-form one-step-ahead forecast errors. The identification of structural shocks in the context of a VAR requires auxiliary assumptions. Ideally, these auxiliary assumptions should satisfy the following coherency requirement: Supposing the identified VAR is fitted to artificial data from the DSGE model, then the VAR estimates of the structural shocks should coincide with the shocks that were fed into the DSGE model to generate the data.

We find that the DSGE-VAR procedure delivers reasonably robust answers to the question, How good is my DSGE model?

The marginal data density analysis as well as the pseudo-out-of-sample forecasting exercise suggests that an unrestricted VAR does not master the first challenge: It fits and forecasts worse than the DSGE model and hence does not provide a credible benchmark.

On the other hand, the DSGE-VAR($\hat{\lambda}$) passes the first hurdle easily. Our procedure of selecting the hyperparameter ensures that we are using a benchmark specification that fits better than the DSGE model.

The second challenge lies in the derivation of a model-consistent VAR identification scheme. For instance, Altig et al. (2004) consider responses to two shocks: a permanent technology shock and a monetary policy shock. The authors identify technology shocks by assuming that these are the only shocks that can have a permanent effect on the long-run level of the real variables (output, consumption, etc.), as in the DSGE model. Monetary policy shocks are identified using the assumption that firms and households can observe them only with a one-period lag. Hence prices, output, and other macroeconomic quantities do not react instantaneously to monetary policy shocks.

It is well known, however, that in short samples long-run restrictions lead to imprecise estimates of impulse responses (see, for instance, Faust and Leeper 1997), making them a possibly unreliable benchmark. Monetary policy impulse responses, identified with short-run restrictions, do not suffer from this drawback. However, for the identification scheme to be model consistent, one typically has to introduce fairly ad hoc decision lags into the structural model. Finally, monetary policy and technology shocks combined explain only a fraction of the variation observed in the data. Yet it is typically not straightforward to construct model-consistent identification schemes for remaining shocks using traditional zero restrictions.

It turns out that the DSGE-VAR framework is rich enough to deliver an elegant solution to this identification problem, as originally discussed in Del Negro and Schorfheide (2004). Recall that a VAR is identified when there is a unique mapping between the forecast errors and economically interpretable shocks. In the DSGE-VAR procedure the mapping is chosen such that the DSGE and the DSGE-VAR impulse responses would coincide if the data were generated by the DSGE model ($\lambda = \infty$). By construction, the identification in the DSGE-VAR is therefore consistent with that in the DSGE model. Whenever $\hat{\lambda}$ is less than infinity, the impulse responses of DSGE-VAR($\hat{\lambda}$) and the DSGE model will differ. From the comparison between the two, one

can potentially learn in which dimensions the DSGE model is failing and how it can be improved. Although we do not discuss the empirical findings in Del Negro et al. (2004) in detail here, we note that the impulse response comparison confirms the results discussed so far. For the No Habit model there is clear evidence that something is amiss: For instance, consumption responds abruptly to both monetary policy and technology shocks for the DSGE model, while the response according to the reference model is smoother. Again, such strong evidence is absent in the case of indexation.

Conclusion

The article discusses DSGE-VAR—a procedure that can be used to evaluate and compare DSGE models. Drawing on existing work by Del Negro et al. (2004), the article also provides examples of how the procedure works in practice. We find that the DSGE-VAR procedure, unlike some of the current practices in the literature, delivers reasonably robust answers to the question, How good is my DSGE model?

REFERENCES

Altig, David, Lawrence J. Christiano, Martin Eichenbaum, and Jesper Linde. 2004. Firm-specific capital, nominal rigidities, and the business cycle. Federal Reserve Bank of Cleveland Working Paper 04-16, December.

Calvo, Guillermo. 1983. Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics* 12, no. 3:383–98.

Christiano, Lawrence J., Martin Eichenbaum, and Charles Evans. 2005. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113, no. 1:1–45.

DeJong, David N., Beth F. Ingram, and Charles H. Whiteman. 2000. A Bayesian approach to dynamic macroeconomics. *Journal of Econometrics* 98, no. 2:203–23.

Del Negro, Marco, and Frank Schorfheide. 2004. Priors from general equilibrium models for VARs. *International Economic Review* 45, no. 2:643–73.

Del Negro, Marco, Frank Schorfheide, Frank Smets, and Raf Wouters. 2004. On the fit and forecasting performance of New Keynesian models. Federal Reserve Bank of Atlanta Working Paper 2004-37, December.

Doan, Thomas, Robert Litterman, and Christopher Sims. 1984. Forecasting and conditional projections using realistic prior distributions. *Econometric Reviews* 3:1–100.

Faust, Jon, and Eric M. Leeper. 1997. When do long-run identifying restrictions give reliable results? *Journal of Business & Economic Statistics* 15:345–53.

Fernández-Villaverde, Jesús, and Juan Francisco Rubio-Ramírez. 2004. Comparing dynamic equilibrium models to data: A Bayesian approach. *Journal of Econometrics* 123, no. 1:153–87.

Haavelmo, Trygve. 1944. The probability approach in econometrics. *Econometrica* 12 (supplement, July): iii–vi+1–115.

Kydland, Finn E., and Edward C. Prescott. 1982. Time to build and aggregate fluctuations. *Econometrica* 50, no. 6:1345–70.

Nason, James M., and Timothy Cogley. 1994. Testing the implications of long-run neutrality for monetary business cycle models. *Journal of Applied Econometrics* 9 (supplement, December): S37–S70.

Otrok, Christopher. 2001. On measuring the welfare cost of business cycles. *Journal of Monetary Economics* 47, no. 1:61–92.

Rotemberg, Julio, and Michael Woodford. 1997. An optimization-based econometric framework for the evaluation of monetary policy. *NBER Macroeconomics Annual* 12:297–46.

Schorfheide, Frank. 2000. Loss function-based evaluation of DSGE models. *Journal of Applied Econometrics* 15, no. 6:645–70.

Sims, Christopher A. 1980. Macroeconomics and reality. *Econometrica* 48, no. 1:1–48.

Smets, Frank, and Raf Wouters. 2003. An estimated stochastic dynamic general equilibrium model of the euro area. *Journal of the European Economic Association* 1, no. 5:1123–75.

Stock, James H., and Mark W. Watson. 2002. Has the business cycle changed and why? NBER Working Paper No. 9127, August.