Posterior Predictive Analysis for Evaluating DSGE Models

Jon Faust
Johns Hopkins University

Abhishek Gupta[*]
Gettysburg College

October 30, 2010

Abstract

In this paper, we develop and apply certain tools to evaluate the strengths and weaknesses of dynamic stochastic general equilibrium (DSGE) models. In particular, this paper makes three contributions: One, it argues the need for such tools to evaluate the usefulness of the these models; two, it defines these tools which take the form of prior and particularly posterior predictive analysis and provides illustrations; and three, it provides a justification for the use of these tools in the DSGE context in defense against the standard criticisms for the use of these tools.

JEL Classification: C11, C52, E1
Keywords: Prior and posterior predictive analysis; Model Evaluation; DSGE.

[*] Corresponding Author: Abhishek Gupta- 300 North Washington Street, Department of Economics, Gettysburg College, Gettysburg, PA, 17325; e-mail: agupta@gettysburg.edu. We thank Frank Smets and Rafael Wouters for sharing their code and estimation results. ©2010 by Jon Faust and Abhishek Gupta. All rights reserved. Short sections of text may be quoted without explicit permission provided that full credit is given to the source.

1 Introduction

Dynamic stochastic general equilibrium (DSGE) models have come a long way. After Kydland and Prescott (1982) demonstrated that a small DSGE model could match a few simple features of the macro dataset, there ensued a 20-year research program of adding both complexity to the model and data features to account for, and the models gradually began to approximate the richness of the macroeconomy.

A watershed event came in 2003 when Smets and Wouters (2003) demonstrated that the family of DSGE models had reached the point that it 'fit' seven key variables about as well as some conventional benchmarks. They explain their main contribution:

> [Our results] suggests that the current generation of DSGE models with sticky prices and wages is sufficiently rich to capture the time-series properties of the data, as long as a sufficient number of structural shocks is considered. These models can therefore provide a useful tool for monetary policy analysis in an empirically plausible setup. (2003, p.1125)

Since Smets and Wouter's demonstration, central banks around the world have rapidly been building DSGE models and pressing them into the service of monetary policymaking.

Not everyone agrees, however, that fitting a few variables as well as standard benchmarks is a sufficient condition for the models to be ready for practical policy work at central banks. For example, Sims (1980) famous critique of the models of the 1970s was more or less that the 70s models were highly problematic *despite* their impressive fit. Further, while the current DSGE models are large-scale models by standards of what is possible to solve and manipulate, they are unquestionably small relative to what would be ideal—standard models do not break out consumer durables, inventories, or housing, and have trivial financial sectors.

Perhaps surprisingly, the debate over the adequacy of DSGE models recently rose to the level of a Congressional hearing. Solow (2010) argued that the models were deeply deficient:

> The national—not to mention the world—economy is unbelievably complicated, and its nature is usually changing underneath us. So there is no chance that anyone will ever get it quite right, once and for all. Economic theory is always and inevitably too simple; that can not be helped. But it is all the more important to keep pointing out foolishness wherever it appears. Especially when it comes to matters as important as macroeconomics, a mainstream economist like me insists that every proposition must pass the smell test: does this really make sense? I do not think that the currently popular DSGE models pass the smell test. (p.1)

Chari (2010), (p.7), agreed that “We do not fully understand the sources of the various shocks that buffet the economy over the business cycle,” but testified that “[Policy advice from DSGE models] is one ingredient, and a very useful ingredient, in policy making.”

Of course, these views reflect a longstanding schism in macro. We seek to avoid this argument over whether the glass is essentially empty or nearly full. To push this tired metaphor to the breaking point, we argue that rather than focusing on how full the glass is, we should focus on just what liquid we are being asked to swallow.

That is, we should focus on what are the particular strengths and weaknesses of the models as they pertain to the intended task—monetary policymaking. Like Tiao and Xu (1993), (p.640), we argue for “...development of diagnostic tools with a greater emphasis on assessing the usefulness of an assumed model for specific purposes at hand rather than on whether the model is true.”[1]

We believe that this perspective is particularly important when discussing models for active use in the monetary policymaking process. Policy must be made. In lieu of a formal, coherent model of important general equilibrium effects, the policymaking process employs an implicit model consisting of sector experts and general equilibrium experts hashing out the issues in (long and painful) meetings. The important issue is

[1] see also, e.g., Hansen (2005).

not some overall quality judgement about the models, but assessing in what ways the models could render the current policymaking process more reliable.

No one, we believe, is asserting DSGE models have reached the point that policy can be placed more or less on model-based autopilot. Thus, we need tools for assessing how DSGE models can best be used to inform expert judgement in the policymaking process.

This paper makes three claims:

First, while DSGE modelling has come a long way, there remains room for improvement in areas that are materially important for policymaking. We think this claim should be entirely uncontroversial but documenting some particulars provides a basis for the later illustration of our inference tools.

Second, prior and particularly posterior predictive analysis can be valuable tools in assessing strengths and weaknesses of the DSGE models. Predictive analysis was originally popularized by Box (1980) and has been extended in many ways. Our contribution is to adapt these tools to the DSGE context and illustrate their usefulness. Most notably, perhaps, we adapt the discrepancy analysis of Gelman et al. (1996), which seems to have gotten little use in macro, to analyze the causal channels in DSGE models.

While prior and posterior predictive analysis have been widely used in many areas, these techniques have also been criticized as being inconsistent with coherent inference. Our third goal is to argue that standard criticisms of prior and posterior predictive analysis, whatever their merits in other contexts, miss the point in the DSGE context. In particular, posterior predictive analysis can be viewed as a natural pragmatic Bayesian response to a murky modelling problem.

These three points come in the next three sections. Along the way we discuss and illustrate various aspects of implementing prior and posterior predictive analysis using a version of the iconic Smets Wouters DSGE model. Smets and Wouters have been incredibly gracious in helping us complete this work.

2 Standard Approaches to Bayesian DSGE Modelling for Monetary Policymaking

2.1 The Macro Modelling Problem

We will take largely as given a basic knowledge of the monetary policymaking process and of DSGE modelling for use in the policymaking process. We provide the following sketch. The policymakers of the central bank meet periodically to assess changes to the values of the policy instruments under their control—the policy interest rates and other policy tools. Taking as given the view adopted at the last meeting, a major task at this meeting is to process the information that has arrived since the last meeting.

This processing is complicated by a number of features of the problem. The potentially relevant information set is high dimensional. While there is no consensus on how high dimensional, the ongoing policymaking process closely monitors hundreds of variables.[2] Certain research supports the view that forecasts of relevant variables based on datasets of, say, 70 or more variables outperform others[3] and that the subjective process that focuses even on broader information in some cases does even better (Faust and Wright (2009)). Further, there is a rich set of dynamic feedbacks among the myriad potentially relevant variables. By general consensus, general equilibrium effects involving the expectations of a large set of heterogeneous agents may be central to the policymaking problem.[4]

The existing policy process relies on many more and less formal tools, but is ultimately heavily judgmental. The goal in DSGE modelling is to build a new model to aid in the process of interpreting incoming data, forecasting, and simulation of alternative policies.

[2] For example, see the Federal Reserve's policy meeting briefing materials that are now available at $http : //www.federalreserve.gov/monetarypolicy/fomc_{h}istorical.htm$

[3] e.g., Bernanke et al. (2005a); Bernanke and Boivin (2003); Faust and Wright (2009); Forni et al. (2005); Giannone et al. (2004); Stock and Watson (1999, 2002, 2003, 2005).

[4] For example, policy is often described as expectations management. See, e.g., Bernanke et al. (2005b).

From a modelling standpoint, forecasting does not strictly require a structural model (that is, a model with explicit causal channels). The simulation of likely outcomes under alternative counterfactual policy assumptions does require a structural model, as does the attribution of unexpected movements in incoming data to particular causes. The standard textbook example of the latter is attempting to sort out whether an unexpected rise in GDP growth is primarily due to supply or demand shocks. As these two causes may have different implications for policy, much policy analysis involves this sort of inference.

This modelling problem is made more challenging by the fact that the relevant body of theory provides only limited guidance on short-run and medium-run dynamics. While theory gives us broad guidance on overall dynamics, small adjustment costs, rule of thumb behavior, and similar effects can dominate shorter run dynamics.

Ideally, we need a model with fully articulated causal structure of a very large and complicated system where theory provides limited guidance on important aspects of the dynamics. The final complication is that we have a single historical sample for the process we are modelling, the world macroeconomy. The models will be specified and refined on this familiar dataset. New information arrives only with the passage of time. By general consensus, the historical sample is not large enough to definitively resolve all important issues. The ongoing lack of consensus on basic questions in macro is clear testament to the idea that our only available sample is not resoundingly informative on all relevant policymaking issues.

2.2 Description of DSGE Modelling and the SW Model

The approach in DSGE modelling is to explicitly state the decision problems of groups of agents—e.g., households and firms—including objective functions and constraints. For example, households choose consumption (hence, saving) and labor supply to maximize a utility function. Firms choose investment and labor input to maximize a profit function. Ultimately the solutions these individual optimization problems, when combined with overall resource and adding up constraints, imply the dynamic behavior of the variables modelled. As noted below, given the complexity of the models, we generally end up working with some approximation to the full model-implied dynamics—for example, a log-linear approximation to the deviations from some steady-state implied by the model.

A key feature of this kind of modelling is that agents are forward looking. In the simplest forms of the decision problems, forward looking agents immediately react to news about future conditions, and adjust their behavior much more quickly than is consistent with the macroeconomic data. Thus, 'frictions' are added to the decision problems in order to slow down what would otherwise be excessively jumpy behavior by agents.

Our example model is a version of the SW model. In particular, we use the model described in Smets and Wouters (2007).[5] The model is an extension of a standard closed economy DSGE model with sticky wages and sticky prices, largely based on Christiano et al. (2005). The model explains seven observables, with quantity variables in real, per capita terms: GDP growth, consumption growth, and investment growth, hours worked, inflation, real wage growth, and a short-term nominal interest rate.

The model introduces a rich set of frictions in the decision problems of agents—the Calvo friction for prices and wages, habit formation in consumption, and investment adjustment costs. The seven structural shocks are assumed to follow exogenous, independent persistent (autoregressive of order 1) processes and are interpreted as shocks to overall productivity, investment productivity, a risk premium, government spending, wage mark-up, price mark-up, and the monetary policy interest rate rule.

Because we will focus on consumption as an example later, it is worth going into a bit more detail on the consumption problem. Consumers are assumed to maximize the expected value of a discounted sum of period utility given by,

$$U_t = \left(\frac{1}{1-\sigma_c} (C_t - hC_{t-1})^{1-\sigma_c} \right) exp \left(\frac{\sigma_c - 1}{1+\sigma_l} L_t^{(1+\sigma_l)} \right)$$

[5] The log-linearized equations of the model are provided in appendix A. Readers are referred to Smets and Wouters (2007) for a thorough explanation of the model equations and frictions.

where C_t is consumption at time t, L_t is labor hours at t and H, σ_c and σ_l are scalar parameters. The consumption portion of the utility function has habit persistence parameterized by h and risk aversion parameterized by σ_c, the coefficient of relative risk aversion (CRRA). Larger values of h tend to imply that agents will be more reluctant to change consumption.

2.3 Estimation

Explicitly Bayesian inference methods are the norm in this area. The methods used are, at a general level, a straightforward application of what we will call the plain vanilla Bayesian approach.[6] In a nutshell, this approach requires data, a parameterized model, and a joint prior distribution for the parameters of the model. The model implies a likelihood function for the data, and the model and prior together exhaustively characterize the state of knowledge of the researcher before the new data arrive. The plain vanilla Bayesian scheme tells us how to update the view reflected in the prior density in light of new information that arrives.

Specifically, call the data Y and the model M_θ, with parameter vector $\theta \in \Theta$. In the DSGE case, the likelihood, $L(\theta|Y)$, is implied by the specification of the economic model. Once one has specified the model and processes for the exogenous driving processes, the decision problems of the agents can be solved and this solution implies a likelihood function. Generally for computational reasons we use a log-linear approximation to the exact solution of the model where the approximation is centered on a non-stochastic steady-state of the model. This approximation gives rise to a linear, vector ARMA structure for the dynamics of the model.

The data tend to be taken from the standard macro data set; the sample period tends to be the longest continuous sample for which data are available and for which the structure of the economy was reasonably stable. This latter is a bit of a judgement

[6] Plain vanilla here describes the application of Bayesian principles to a plain vanilla context. This does not tell us what Bayesian principles imply in a richer context such as—we argue—the one described here.

call. SW estimate the model using US data from 1966Q1 to 2004Q4.

Often there are multiple choices to be made for choosing model-based analogs to key quantities in the model. For example, analysts sometimes use only nondurables plus services consumption as the measure of consumption, since the model does not have durables. SW's consumption measure includes durables.

The prior, p_0 is a joint density for θ over the parameter space Θ. The conventional prior used in DSGE modelling varies, but in general terms the formal prior is often specified as a set of marginal distributions for each individual parameter. These are taken to be independent, implying the joint distribution for the prior. Generally, some natural support for each parameter is implied by economic principles, technical stability conditions for the model, and/or earlier applied work. The prior is specified to be fairly dispersed over this support. Through trial and error, the analyst may find regions of the parameter space in which the model seems ill-behaved in some way, and the support is narrowed.

While the papers often have arguments justifying the support of the prior and, perhaps, where it is centered, we have seen no argument that the joint prior implied by the independent specification of marginal priors for the parameters has any justification or has any tendency to produce results that are consistent with any subjective prior beliefs.[7]

Given the state of knowledge reflected in the model+prior and the new information in the sample, a straightforward application of Bayes' law gives us,

$$p_1(\theta) \equiv pr(\theta|Y^r) = \kappa p_0(\theta) L(Y^r|\theta)$$

where κ is the constant that makes the integral of the expression on the right (with respect to θ) integrate to one.

The prior and posterior densities for the two key consumption parameters are shown in Figure 1. Both posteriors are centered at about the same place as the prior, but

[7] In some cases some span of data at the beginning of the available sample is used as a 'training sample' so that the prior is tuned to reflect that dataset. We take this up below.

the posteriors are considerably more peaked, indicating that that data are somewhat informative. The habit persistence parameter is fairly large, suggesting that agents are highly averse to changing consumption; the CRRA parameter is in the range that has become conventional for estimates of models like this on this sample.

2.4 Material Deficiencies, Omissions and Coarse Approximations

Given the size of the system being modelled and the current stage of understanding of the relevant mechanisms and modelling techniques and related algorithms, it remains the case that existing DSGE models involve coarse approximation to some economic mechanisms believed relevant for policymaking and omit other such mechanisms entirely. This is meant to be a description of the current state of development, not a criticism. To motivate the remainder of the paper, it is useful to provide some detail on the state of modelling.

Consider omitted mechanisms and phenomena. Most standard DSGE models do not separately treat durable goods, inventories, or housing, despite conventional wisdom that these items play an important role in business cycles. Many experts believe that credit spreads have important predictive content that might be important for policymaking (e.g., Gilchrist et al. (2009)), but defaultable debt is not modelled. Indeed, until the crisis, the financial sector of these models was entirely trivial. This list of omissions could obviously go much longer.

It is also true that the modelling of phenomena included in the model is often best viewed as a coarse approximation relative to the best knowledge of specialists in the particular area.

For example, important aspects of individual behavior toward risk is parameterized by the coefficient of relative risk aversion. In the best tradition of microfounded modelling, we might ask experts in individual behavior toward risk what values for this parameter might be appropriate. Unfortunately, expert opinion is overwhelmingly clear on one point: individual behavior toward risk is a rich phenomenon not well captured by this single parameter.[8] Many micro phenomena simply cannot be accounted for under this assumption. Suppose we tell the expert we are viewing this as a representative agent approximation to underlying behavior, but would still like guidance on the value. The expert should then remind us that different values will be best depending on the goals of the approximation: to 'fit' the equity premium from 1889 to 1978, CRRA $>$ 10 (Mehra and Prescott (1985)); to 'fit' aggregate lottery revenue probably requires risk loving; to fit the reaction of consumption to changes in monetary policy, probably some value not too far from one is appropriate. As a description of individual behavior, CRRA specification is a crude approximation. The choice of parameter value should be based mainly on how best to center the approximation for a particular purpose.

Analogous issues can be raised about the treatment of habits. At the individual level there is little evidence for strong habits (Dynan (2000)), but strong habits in the model seem to be needed to 'fit' the smooth evolution of aggregate consumption data. Alternative explanations for the aggregate persistence include serially correlated measurement error (Wilcox (1992)) aggregation biases (Attanasio and Weber (1993)), and 'sticky expectations' (Carroll et al. (forthcoming)).

We could do a similar analysis in these models of the labor market, investment, and the financial sector. Our goal is to provide some concrete meaning to the statement that many arguably relevant mechanisms are omitted, many included mechanisms are coarse approximations.

Finally, it also the case that the prior used in these analysis is far from the idealized case in which the model+prior fully reflects the subjective views of the relevant analysts. As Del Negro and Schorfheide (2008) note, there is no reason to suppose that taking reasonable marginal priors for the parameters and treating these as independent will lead to reasonable general equilibrium implications for the model as a whole. As we shall see below, it is not difficult to find examples where the prior is highly informative and at odds with conventional wisdom.

[8] For a good review, see Camerer (1995).

Rather than focusing on particulars, however, our point about that material deficiencies remain is probably best illustrated by the revealed preferences of modelers at central banks. The models remain in a state of substantial ongoing refinement and revision.

3 Prior and Posterior Predictive Analysis

The plain vanilla scheme described above tells us how optimally to shift our views on the relative plausibility of different parameter values $\theta \in \Theta$. But it can never cast doubt on whether the model as whole, $M_\theta, \theta \in \Theta$, is adequate. In any context, this is potentially troubling—George Box famously reminds us, all models are wrong–but it is particularly troubling in a context where we are using an *ad hoc* prior over a model with materially important aspects of approximation error and omission.

Box popularized a family of tools for checking whether an admittedly wrong model might be useful based on prior and posterior predictive analysis. Box's ideas have been elaborated in a number of ways in the statistics and economics literature.[9]

In this section, we take as given the conventional pragmatic arguments in favor of prior and posterior predictive analysis and illustrate the way it could be used to highlight strengths and weaknesses of DSGE models.[10] The basic analytics of prior and posterior predictive analysis are all well-established in the statistics literature. Our contribution is to adapt these tools in ways particularly useful in DSGE work.

3.1 Prior and Posterior Predictive Analysis Defined

Predictive analysis relies on simple idea: if the available sample is too freakish from the standpoint of the model+prior or model+posterior, then perhaps the model or prior

[9] For example, Geweke (2005, 2007, 2010); Bernardo (1999); Gelman et al. (1996); Lancaster (2004).

[10] It might seem more natural to give the theoretical justification before the applications. Our theoretical arguments, however, turn on unique practical aspects of the DSGE context that are best discussed after seeing some concrete examples.

should be refined.[11]

The essence of the argument can be seen in a simple example. You are attempting to do inference on the probability of observing a draw greater than 3 from some random variable. The model states that the sample is independent and identically distributed (iid) draws from a Gaussian with unknown mean and variance one. The prior for the mean is uniform on $[0, 1]$. You obtain a sample of 50 observations and notice that the histogram of these observations is highly skewed to the right (Figure 2, panel (a)). This sample would be very unlikely to arise if the outcomes were indeed iid Gaussian.

The generic idea of predictive analysis is that one might want to reconsider the model+prior at this point, but two additional ideas are worth emphasizing. The right skewness may be materially important to the task at hand, since right skewness could have a large effect on the probability of observing values greater than 3. Thus, one can focus on *relevant* features. Second, one sensible option would be to obtain another sample. But we are focusing on the case in large-scale, general equilibrium macroeconomics—new information will only arrive slowly.

Predictive analysis provides formal tools for judging the degree to which relevant features of a sample are freakish from the standpoint of the model+prior. By *feature*, we mean any well-behaved function of the data: $h(Y)$.[12] Following the spirit in much macroeconomics one might think of these as empirical measures corresponding to some 'stylized fact.' In the example just given, it would be natural to use the sample skewness as the data feature.[13] The sample skewness for the example sample is 1.05, whereas the population skewness of any Gaussian is zero. However, one might wonder how likely one would be to observe a *sample skewness* of 1.05 in a sample of size 50 from the

[11] It might seem most natural to change the model, but in cases like DSGE modelling where the prior is substantially arbitrary, it is not unnatural to think of deciding that the arbitrary choice of prior had put mass in 'the wrong place.'

[12] Box called these *model checking functions.*

[13]

$$h(Y) = \sum_{t=1}^{T}(y_t - \bar{y})^3/(\sum_{t=1}^{T}(y_t - \bar{y})^2)^{3/2}$$

where y_t is the t^{th} observation and $\bar{y}$ is the sample mean.

model+prior at hand.

The model+prior imply a marginal distribution for any $h(Y^{rep})$, where Y^{rep} is a sample of the size at hand drawn according to the model+prior:

$$F_h(c) \equiv \text{pr}(h(Y^{rep}) \leq c) \tag{1}$$

Define Y^r to be the realized sample. One can plot the implied density, $f_h(x)$, along with the realized value $h(Y^r)$ on the sample get a sense of whether the realized value is freakish. Our example model+prior can indeed produce samples with large positive and negative values for the sample skewness, but would do so very rarely—the sample value of 1.05 is far in the tail of the predictive distribution.

Where large values are considered unlikely, Box suggested a prior predictive p-value defined as,

$$1 - F_h(h(Y^r)).$$

This is the probability of observing $h(Y)$ greater than the realized value in repeated sampling if the data were generated by the model+prior. For our example, the p-value is 0.002, or 0.2 percent.

There are, of course, dangers in summarizing a distribution with a single number such as a p-value. Such crude summaries should be used with caution, and we will largely report the entire predictive density. Still at times, p-values provide a convenient and compact summary.

We can use the posterior for the parameters of the model, p_1, instead of p_0 in (1) in computing the predictive density, to obtain the posterior predictive distribution and posterior predictive p-value. Once again, these predictive densities depict the likelihood of observing specified sample features in repeated sampling from the model+prior or model+posterior.

The data features we have discussed so far are a function of Y alone. In modeling causal channels we are not only interested in description, but in why events happen the way they do. To shed light on causal channels, it is also useful to consider features that are a function of the sample and θ: $h(Y, \theta)$. We'll call the former 'descriptive' features and the latter 'structural' features, to emphasize the dependence of the latter on the structural parameter.

Gelman et al. (1996) have written extensively on what we call structural features.[14] These seem to have received little application in macroeconometrics.

Since structural features depend on the unobserved value of θ, they are a bit more subtle to understand than descriptive features. A symptom of the difficulty is that even after observing the sample, there is no single realized value on the sample at hand. However, conditional on any fixed θ^*, we can compute $h(Y^r, \theta^*)$ and, thus,

$$\text{pr}(h(Y^{rep}, \theta^*) > h(Y^r, \theta^*))$$

where Y^{rep} is a random sample of the same size as Y^r drawn according to θ^*. Conditional on θ^*, this corresponds to the $p - value$ computed above. As always we can integrate out the dependence on the unobserved θ^* using the prior or posterior to get, $\text{pr}(h(Y^{rep}, \theta^{rep}) > h(Y^r, \theta^{rep}))$, where θ^{rep} is drawn according to the prior or posterior. This is analogous to the p-value computed above.[15]

Gelman et al. suggest the following computational approach, which may aid in understanding the above expression. Focus on the posterior version for concreteness. The model+prior imply a joint distribution for θ and Y. Thus, we can assess $\text{pr}(h(Y^{rep}, \theta^{rep}) > h(Y^r, \theta^{rep}))$ by repeating the following steps a large number of times, where on the j^{th} step we,

1. Draw $\theta^{(j)}$ according to the posterior

2. Compute $h(Y^r, \theta^{(j)})$

[14] Gelman et al. call $h(Y, \theta)$ a *discrepancy variable* or simply *discrepancy.* The idea is that the feature is meant to help detect a discrepancy between the model and sample.

[15] An alternative for dealing with the unobserved θ is to create some summary scalar. We could examine the value at the posterior mode, or the mean value for the feature where the mean is taken with respect to the prior or posterior. We discuss how our approach complements these others in the applications below.

3. Draw $Y^{(j)}$ according to $\theta^{(j)}$

4. Compute $h(Y^{(j)}, \theta^{(j)})$

5. Save the pair $h(Y^r, \theta^{(j)}$,$h(Y^{(j)}, \theta^{(j)})$

The marginal distribution of the $h(Y^r, \theta^{(j)})$s is a density corresponding to the realized value in descriptive features. The marginal distribution of the $h(Y^r, \theta^{(j)})$s is the posterior predictive distribution for this feature.

The scatter plot of $h(Y^{(j)}, \theta^{(j)})$ (on the vertical axis) against $h(Y^r, \theta^{(j)})$ (horizontal) will give a sense of the joint distribution of the two items, and the share of points falling above the 45 degree line is an estimate of the p-value described above. This p-value is the probability in repeated sampling from the model+posterior, that we observe a sample generated under a θ_0 for which the h exceeds the h implied by θ_0 on the realized sample.

Obviously, if it is small values that one wishes to detect then, the share of points under the 45 degree line constitutes a p-value. More generally, inspection of the joint distribution will, once again, be more informative than a simple p-value computation.

Note that the predictive p-value for descriptive statistics can be computed using a simplified version of the same algorithm exploiting the fact that h does not depend on θ. Thus, the second step above may be computed outside the loop and the realized density collapses to a point and all we have to plot is the marginal predictive density and the realized point value. These are the algorithms we use in the examples reported below.

3.2 Illustrations I: Descriptive Features

In this section, we illustrate how these techniques can be used to discover and highlight strengths and weaknesses of DSGE models using the SW model. This is not intended as a thorough substantive critique of this model; rather, we present examples meant to illustrate the functionality of the methods. We attempt to provide a substantive analysis in other papers (Gupta (2010), Faust and Gupta (2010b), Faust (2009)).

It is useful to keep in mind two forms of analysis that are complementary to what we are advocating: moment matching and full blown likelihood analysis. In traditional moment matching with a DSGE model, one selects (either by estimation or calibration) values for the parameter, θ, and then compares population moments implied by the model to the corresponding sample moments for the sample at hand. In a full-blown Bayesian-inspired likelihood analysis, the emphasis is on comparing models or parameter values based on the relative likelihoods, perhaps, as weighted with the prior. We seek to emphasize how posterior predictive analysis can be a complement to each.

In traditional moment matching as started in the DSGE literature by Kydland and Prescott (1982), one might focus on the some version of the variance covariance matrix of the variables—say standard deviations, correlations, and autocorrelations. In SW for example, the correlation of inflation and output growth at the posterior mode is -0.22, while the corresponding sample correlation is -0.31. It is difficult to assess whether this is a success or failure of the model, in part, because this comparison fails to represent two potential areas of uncertainty. First, summarizing the model only by the correlation at a single θ does not reflect uncertainty in the choice of parameter. Replacing the single value at the posterior mode with the posterior density for θ will bring the uncertainty in θ into the comparison (Figure 3, blue dashed line). To emphasize, the blue line is the posterior density for the population correlation implied by θ. Since the sample value is relatively far into the tail of the posterior density, one might once again take this as evidence against the model.

There is a second aspect of uncertainty, however: the sample correlation is not a precise estimate of the population correlation in the underlying process driving the economy. Regardless of the population correlation, the model+prior could, in principle, imply that any sample correlation might be observed in a small sample.

The posterior predictive density tells us what sort of values we would expect to see for the *sample correlation* in repeated sampling with sample size equal to the sample size at hand. It turns out (Figure 3, black line) that under the model+posterior, there is nothing particularly freakish about seeing *sample* correlations of -0.31 when the *population* value is -0.22 at the posterior mode.

In this case, it is important to keep the interpretation in mind: the model is consistent with the data essentially because the model implies that the sample correlation will be poorly measured—correlations like that observed in the data are likely to be observed even when the true correlation is quite different.

Posterior predictive analysis provides a way to investigate and highlight known problem areas of the model. For example, the correlation of consumption and investment growth in DSGE models has been a continuing problem. For most countries consumption and investment growth are strongly positively correlated: booms and busts tend to involve both consumption and investment. There are forces in the model, however, that tend to drive this correlation toward zero.[16] In the SW model, the posterior for this correlation is centered on low values (Figure 4, panel(a), black solid line), and the sample value over 0.5 is far in the tail of the posterior. In this case, the posterior predictive density is slightly more dispersed (Figure 4, panel(b), black solid line), but the *p*-value remains below 1 percent. Formally, if samples were repeatedly drawn from the model+posterior, less than 1 percent of draws would give values as extreme as that observed on the sample.

In cases where the model+posterior suggest that that the sample at hand is freakish, there are three natural diagnoses: i) strange samples happen, ii) the model needs refinement, and iii) the prior needs refinement. This last possibility, of course, arises particularly when the prior has *ad hoc* elements. To shed some light on this latter possibility we can look at the prior predictive density. If the prior predictive density strongly favors low or negative correlations of investment and consumption growth, then the posterior result could be due to the unfortunate choice of prior. In the current example, however, this is not the case (Figure 4, panel(b), blue dashed line). The marginal prior for this sample correlation actually favors large positive correlations.

For the SW model, the update using the data overpowered the strong prior and

[16] For example, a productivity shock that raises real interest rates may raise investment but reduce consumption due to the increased incentive to save.

pushed the posterior estimate to the far side of the sample value. This illustrates the complexity of working with large dynamic systems. The θs that give large positive correlations of consumption and investment must have been downweighted by the likelihood because those θs have some other implication that is at odds with the sample. Smets and Wouters included what they call the risk premium shock in their model with the express purpose of boosting the correlation of consumption and investment growth. This shock seems to do the trick in the prior, but not the posterior. We return to this structural issue below.

These examples were intended to illustrate how posterior predictive analysis could complement or extend the sort of moment matching exercises that have been common in the literature.

Of course, defenders of full-blown Bayesian analysis have long criticized moment matching. Looking at a few marginal distributions for individual moments is no substitute for a metric on the whole system, and the likelihood itself is the natural way of summarizing the full implications of the model. Full likelihood analysis may show, for example, that posterior odds favor DSGE model A over DSGE model B. In the analysis suggested by Del Negro et al. (2007), one forms a Bayesian comparison of the DSGE model to a general time series model. In this case, one can learn that data shift posterior plausibility mass along a continuum from the fully articulated structural model to the general model with no causal interpretability. This sort of comparison may be very useful as an overall metric on how the model is doing.

We are considering model building for ongoing, real-time policy analysis, however. All the models have material deficiencies and are under ongoing substantial revision. Thus, echoing our second and third main points from the introduction, we argue that in addition to full-blown likelihood analysis it is important systematically to explore the particular strengths and weaknesses of the model salient to the purpose of policymaking. The fact that the posterior shifts mass from one model toward another is not very revealing of the particular strengths and weaknesses of either. We argue that by using a richer set of data features than simple moments, some important aspects of the models can be revealed.

For example, Gupta (2010) argues that an important part of policy analysis at central banks is interpreting surprising movements in the data. Policy at one meeting is set based on anticipated outcomes for the economy. At each successive meeting, policymakers assess how new information has changed the outlook and what this implies for the appropriate stance of policy. In a formal model, this amounts to interpreting the one-step (where a step is one decision making period) ahead forecast errors from the model.

The simplest substantive example of this perspective comes in the textbook aggregate supply/aggregate demand model. If prices and output come in above expectation, one deduces that an unexpected positive shock has shifted AD. If output is higher but prices lower than expected, one deduces that a favorable supply shock has shifted aggregate supply. In the textbook case, the two outcomes have different policy implications.

The key insight Gupta argues for is that policymakers need more than a model that forecasts *well* in some general sense. They need a model that properly captures the joint stochastic structure of the forecast errors. As a simple way to examine the properties of the model in this regard, we can take our descriptive feature to be the correlation of one-step forecast errors out of a benchmark time series model estimated on the sample. For example, one could use a first order vector autoregression (VAR), a Bayesian VAR, or a VAR with lag length set by the AIC between 0 and 6. All that is required is that based on the sample alone, one can evaluate the value of the feature. For our illustration example, we use the simplest of these, the first-order VAR.

Since output growth is a major focus of policy, we focus for this example on the correlation of the output growth forecast error with the errors for the other 6 variables. The prior and posterior predictive densities along with the sample value are examined in Figure 5.

Several notable results can be seen. First, the prior is highly opinionated putting most mass on correlations near one. This illustrates again that although the marginal prior distributions for the parameters are fairly dispersed, the joint implications of the largely *ad hoc* prior for questions of interest in policymaking may be highly concentrated. Indeed, it may be highly concentrated in regions that do not reflect any subjective prior judgements. No expert believes that if we could just forecast output growth properly we could also nail a forecast for hours of work, but this view is reflected in the mass near one in Figure 5, panel(c) (blue dashed line).[17]

The realized value of the forecast error correlation is fairly generally far in the tail of the posterior predictive distribution. In particular, the relation between output and inflation (two key policy variables) is problematic. Of course, the literal meaning is that the sample is freakish from the standpoint of the model+posterior. More provocatively, in practice on the realized sample, policymakers were systematically faced with the problem of interpreting inflation and output growth surprises of opposite signs (negative realized correlation in Figure 5, panel(d)). The model+posterior says that this pattern was a freak outcome and policymakers need not worry much about facing this problem in the future.

This simple example is only illustrative. Policymakers will in practice use a more sophisticated forecasting model. Thus, one might ideally choose a more sophisticated benchmark forecasting model. Or one could analyze the properties of optimal model-consistent forecast errors. Gupta (2010) provides a version of this more complete analysis.

3.3 Illustrations II: Structural Features

Ultimately, policymakers must go beyond the descriptive in order to draw inferences about the causes of economic variation and the likely causes of policy responses. Identifying causal structure in macro is very contentious, and when using a large model, the problem is multiplied by the complexity of the system. It is very difficult to look at a model and judge whether the causal structure as a whole is broadly consistent with

[17] A complete diagnosis of this fact is beyond the scope here. However, it appears that this is due to the fact that all the shocks enter the prior with the same parameters. Despite being ‘the same’ in this nominal sense, a given variance shock means something different economically depending on how it enters the model. In this case the result seems to be that the prior is that demand shocks dominate.

any given view. One natural way to focus the examination is to analyze what 'causal story' the model tells of the fluctuations in the familiar sample. In doing so, we shift the large and amorphous question, 'how does the model say the world works?' to 'What light does the model shed on the sample that is the source of our current expertise and conventional wisdom?'

In the current literature it is common to present a historical decomposition of headline variables like GDP growth in terms of the underlying structural shocks. Technically, for any value of θ, we can compute our best estimate of the underlying latent structural shocks. Given the linear Gaussian structure assumed for the model, these can be computed with the Kalman smoother[18]. The standard practice seems to be to produce a historical decomposition in terms of the smoothed shocks evaluated at the posterior mode for the parameter. For example, (Figure 6) taken from Smets and Wouters (2007) shows that the SW model attributes much of the deep recession in 1982 to the collective effect of demand shocks in the model. Demand shocks also account for much of the recession in 2001.

These decompositions can be a very useful tool for understanding the models, and posterior predictive analysis of structural features can form a valuable complement to these historical decompositions. First, note that making judgments about the model based on decompositions like this has the same problems that arise in the simple moment matching discussed above: it ignores uncertainty in θ and if something seems amiss, it provides no systematic way to judge just how implausible or freakish the result is.

There are many ways to use posterior predictive analysis to provide more systematic results complementary to the historical decompositions. For example, for a broad overall check we can take our structural feature, our $h(Y, \theta)$, to be elements of the sample correlation matrix of the smoothed estimates of the structural shocks. Remember that the structural shocks are assumed to be mutually uncorrelated in the model.

Each θ (when combined with a sample) implies a sample correlation matrix for smoothed shocks, so this is a well-defined structural feature. Since θ is unknown, we will

[18] Harvey (1991)

not have a single realized value on the sample at hand; instead we will have a posterior density for the realized value that takes into account our remaining uncertainty about θ. Following Gelman's suggestion, we can, however, consider the joint posterior density for the feature on the realized sample and predictive samples.

For example, take as our structural feature the sample correlation of the risk premium shock and the price markup shock. We represent the posterior predictive information (Figure 7, panel(f)) in a scatter plot in which each point represents a joint draw of a θ and a replication sample. For each such draw we compute and plot the pair $(h(Y^r, \theta), h(Y^{rep}, \theta))$—the feature on the realized and on the predictive sample, respectively. We plot these pairs as a scatter plot with the realized value on the horizontal axis and predictive sample on the vertical axis. If a typical draw from the model+posterior implies a sample correlation like that implied for the realized sample, the points of the scatter will lie around the 45 degree line.

For the chosen correlation, the entire point cloud is well below the 45 degree line. This implies that there is no value for the structural parameter that is likely to produce a correlation of these two structural shocks as high as that implied on the realized sample.[19] (Note: in scatter plots like Figure 7, the number in the upper left corner of the panel is the share of points on whichever side of the 45 degree line has a smaller share of points.)

The point cloud in this case is tall and thin. The fact that the point cloud is quite narrow tells us that for the bulk of θs getting posterior mass, the sample correlation on the realized sample is a bit over 0.2: for all relevant θs, the sample correlation of these shocks was substantial and in the 0.2 range. The height of the cloud spans values from about -0.15 to +0.15 and is roughly centered on zero. This gives a sense of what we would expect to see for this sample correlation in repeated sampling. Values vary some, but are mainly clustered near the population value of zero. There is very little chance that the model+posterior would generate a sample implying a sample correlation between these shocks as high at 0.2.

[19] Actually, we should say there is no θ getting nontrivial posterior mass with the stated property.

There is no sign of discrepancy between the data and model+posterior for the sample standard deviation of the risk premium shock (Figure 7, panel(a)) or its correlation with the wage markup shock (Figure 7, panel(g)). The correlation with the productivity (Figure 7, panel(b)) and government spending shocks (Figure 7, panel(c)), however, on the sample are far from what the model+posterior would be likely to generate.

What is the interpretation? Of course, nature may have given us a sample that just happened to imply large sample correlations among the smoothed shocks. If we set this possibility aside, we must consider misspecification of the causal channels in the model.[20] The logic of the model is that these shocks originate in behaviorally distinct sectors of the economy and the population correlation is zero in the model (for every θ). In order to accommodate the sample using the causal channels specified in the model, however, various of these causal forces had to systematically work in concert. If one strongly believes that these forces are originating in behaviorally distinct sectors of the economy (as is the standard assumption) then the model needs refining. Otherwise, some causal account of the linkages must be specified.

As noted above, the risk premium shock was included in the model to help accommodate the positive correlation of consumption and investment growth in the sample. Above, we saw that difficulties remain in this descriptive feature. This analysis provides further evidence that perhaps that aspect of the model is misspecified. Gupta (2010) argues that a more systematic look at the full set of correlations among the structural shocks can provide valuable clues to economic sources of the misspecification and guide future model refinement.

There are many ways to go beyond mere correlations of structural shocks to focus on issues important in policymaking. For example, understanding and avoiding inefficient recessions is one goal of monetary policy. Thus, it is natural to focus on how the model helps us understand the recessions in the sample.

In the U.S., it is well known that periods of at least two consecutive quarters of

[20] In principle, the result could be an artifact of the prior once again, but the fact that the population correlation is zero for every θ and a check of the prior predictive distributions suggests this is unlikely.

negative GDP growth correspond fairly closely to the NBER's definition of recessions. Thus, on any sample, we can partition the observations of the smoothed structural shocks into those occurring in an episode of at least two quarters of negative GDP growth and the others. We can examine the posterior predictive description the model provides of the recessions in the sample.

For example, take as our feature the sample standard deviation of the smoothed risk premium shock during periods of recession and boom (Figure 8). The analysis provides no indication of problem with the standard deviation on the full sample or during booms (Figure 8, panel(a) and panel(c)). The sample standard deviation during recessions on the realized sample, however, is much higher than we would expect to observe under the model+posterior (Figure 8, panel(b)). Indeed, we can take as our feature the difference in the sample standard deviation in recession and boom periods. This difference (Figure 8, panel(d)) is once again considerably larger than we would expect to see out of a sample drawn at random from the sample+posterior.

Recessions in the post-War sample, according to the model, were a collective freak occurrence of abnormally large risk premium shocks occurring systematically at business cycle frequencies. Faust and Gupta (2010b) provide a more complete analysis of this topic finding similar results for other structural shocks. Once again, we can accept that nature gave us a very strange sample in which some of the focal events for policymakers—recessions—repeatedly arose for reasons that are highly unlikely ever to be repeated, or one can consider model+prior refinement.

The main point in this section is to illustrate potential uses for posterior predictive analysis, and not to provide a substantive analysis of the SW model. Indeed, there are many thorough critiques of this model, including very incisive critique by the authors themselves, especially in joint work with Del Negro and Schorfheide (Del Negro et al. (2007)). We argue that the sort of posterior predictive analysis illustrated above provides a complementary tool and is especially useful for highlighting strengths and weaknesses of the models as they pertain to particular uses such as policymaking.

3.4 Elaborations and Abuses

We have deliberately stuck to fairly straightforward illustrations in the previous section in order to introduce these tools. There are many natural elaborations. For example, as Gelman et al. note, one might want to condition all the computations on certain data features. We have deliberately focused on basic applications of the ideas behind prior and posterior predictive analysis in order to introduce the ideas.

We have examined many different data features. Of course, whenever one has multiple statistics there are multiple ways they might be combined and consolidated. For example, one could take account of the full joint distribution of some group of features. Thus, one could ask how likely the model would be to produce sample jointly showing values as extreme as the realized values. One could also combine the features into some *portmanteau feature* and only consider the marginal distribution of the overall combined feature. We have argued for the benefits of interpretability of the features and such a portmanteau would probably lose some of that.

Any discussion of the myriad features one might combine naturally leads to a discussion of how this approach might be misused and abused. On any sample, we can define features that are as 'freakish' as we like in the sense of being present in a small proportion of all samples. Indiscriminate assessment of long lists of features will, with probability one, lead one to discover that each random sample (like each child) is special in its own way. As Gelman et al. (1996), Hill (1996) and many others emphasize, any tool like this must be used with judgement.

In particular, we are suggesting using these tools to highlight areas of consonance and of dissonance between the model any strongly held views about the only existing sample. As we raise this topic, however, we begin to squarely face the third major point of the paper: the standard criticisms of prior and posterior predictive analysis and what we argue is a nonstandard defense in the DSGE context.

4 Standard Criticisms and a Nonstandard Defense of Posterior Predictive Analysis

While we believe that posterior predictive analysis could be an extremely useful tool, uses so far in the DSGE literature have been limited. This may, in part, be because of the strong arguments often stated against the coherence of this form of inference.

In this section, we argue that standard arguments against posterior predictive analysis—whatever their merits in other cases—are moot or miss the point in the DSGE context. Indeed, posterior predictive analysis is arguably a natural pragmatic attempt to apply Bayesian principles under challenging conditions.

4.1 Standard Criticisms

Prior and posterior predictive analysis like all inference based on hypothetical other samples, violate the likelihood principle, which "essentially states that all evidence, which is obtained from an experiment, about an unknown quantity θ, is contained in the likelihood function of θ for the given data..." (Berger and Wolpert, 1984, p.1). Given that prior and posterior predictive analysis share with frequentist analysis an emphasis on behavior in repeated sampling, many of the objections echo the arguments in the familiar frequentist vs. Bayesian debate.

One cannot deny that troubling problems can arise whenever one attempts to cast doubt on a model when based on the fact that was observed in the sample was less likely than other samples that were not in fact observed.[21] One way of seeing the problem of casting doubt on a model due to the fact that the sample at hand is unlikely is that this approach begs the question 'unlikely compared to what?' We have no alternative model that renders the existing sample more likely. Inference without an explicit alternative is fraught a host of problems, leading some to the summary judgement (Bayarri and Berger, 1999, p.72), "The notion of testing whether or not data are compatible

[21] The literature here is immense. For a recent treatment aimed at economists, see Geweke (2010).

with a given model without specifying any alternative is indeed very attractive, but, unfortunately, it seems to be beyond reach."

Geweke (2010), (p.25), argues that, while both prior and posterior predictive analysis violate the likelihood principle, but posterior predictive analysis involves "a violation of the likelihood principle that many Bayesians regard as egregious." This is in part because in this analysis *uses the data twice* in an important sense (without taking formal account of this fact). One checks the freakishness of the sample using the posterior that was already updated using that same sample.

Berger and Wolpert (1984) offer the following judgement about the use of such techniques:

> Of course, even this use of significance testing [as proposed by Box] as an alert could be questioned, because of the matter of averaging over unobserved x. It is hard to see what else could be done with [the maintained model] alone, however, and it is sometimes argued that time constraints preclude consideration of alternatives. This may occasionally be true, but is probably fairly rare. Even cursory consideration of alternatives and a few rough likelihood ratio calculations will tend to give substantially more insight than will a significance level, and will usually not be much more difficult than sensibly choosing T [the data feature] and calculating the significance level. (p.109)

We begin our defense of posterior predictive analysis by accepting (or, at least, choosing not to contest) essentially all of Berger and Wolpert's points. In particular, until recently, Berger and Wolpert's claim that specifying explicit alternatives, perhaps in a cursory manner, is easy was nearly a tautology. Until recently, Bayesian methods were computationally infeasible except in trivial cases and constructing cursory alternatives in such cases may be easy. Constructing meaningful alternative models of the world macroeconomy with fully articulated causal channels, we argue, is not easy. And, thus, we claim this is one of the rare cases.

We also agree with Berger and Wolpert that when one comes across a rare cases where specifying alternatives is not trivial, it is difficult to imagine any systematic way to proceed other than some variant of the basic idea laid out by Box, and that is what we are proposing. But our defense of posterior predictive analysis goes considerably deeper based on a number of features that distinguish the DSGE context from that contemplated in the plain vanilla scheme.

4.2 Nonstandard Defense of Posterior Predictive Analysis for a Nonstandard Context

One slowly growing sample. In DSGE modeling, macroeconomists are attempting to formalize and reify our understanding of the world economy. Unfortunately, while the available sample regarding the general equilibrium process is growing, it grows sufficiently slowly that we may treat it as fixed.[22] The many rounds of refinement take place using the same data and the formal prior and posterior in all DSGE work are best viewed as two different ways of using the same sample. The 'posterior' for the current model will soon be replaced by another 'posterior' for a revised model, and this new 'posterior' will be computed on the same data as were the many previous versions.

For the remainder of this section by *posterior* in italics we simply mean what this term has come to mean in the DSGE literature: the result of the latest round of update on the familiar sample.

Given the intertwining of expertise and the only sample it is almost inevitable that expert have certain strongly held prior views regard the current sample. For example, at the most general level, most macroeconomists believe that systematically repeated features of the business cycle are in fact systematic features of the underlying mechanism. Posterior predictive analysis then can be seen as a pragmatic way to check the consistency of the model+*posterior* with difficult to impose prior views.

[22] by general consensus (confirmed over the 30 year development process) the new information in a few observations is unlikely to provide much additional information. Observations like those from the recent crisis are arguably highly informative, but generally raise more questions than answers.

The current model by general consensus remains materially deficient. The model +*posterior* , whatever its merits, will be used to inform current decision making and as a basis for the next round of model refinement. Much of the analysis we advocate might, in principle be carried out using prior predictive analysis and Geweke (2010), (p.24), makes a strong case for doing so based on the position that we should place "specification analysis ahead of formal inference." Prior predictive analysis can be used in this specification analysis step, but any posterior analysis on what will be treated as the posterior update sample must be confined to the formal inference step.

This analysis does not conform well to the case where we have ongoing refinement of a materially deficient model that is going on in parallel with actual decision making based on the current best model. Substantive ongoing specification analysis has in the past and will for the foreseeable future accompany use of the current *posterior* . Perhaps the simplest way to avoid this debate is to consider the current *posterior* to be the prior for use in the next stage of ongoing specification analysis.

As an overarching idea, however, we think it is uncontestable that if the current model+*posterior* are thought to be materially flawed and yet will be used in policy analysis, searching for problematic predictions from this model+*posterior* must be consistent with good sense as well as Bayesian principles.

There is an alternative to the model—the current subjective policymaking machinery. In the context of DSGE modelling for policy analysis, it is important to emphasize there is an alternative: the implicit model in the current subjective policymaking process. A natural inclination might be to argue that we should render that model formal and explicit, and that this newly formalized model is the model to which we should compare the DSGE model. Of course, this misses the point entirely.

The current DSGE models built for policymaking *are* the current state of our efforts to reify the current policymaking process—to formalize the good and throw out the bad. Our defense of posterior predictive analysis is that it can render the process of enriching and refining the model more efficient by focusing attention on areas that are most troubling from the standpoint of the task at hand.

4.3 An Alternative: Patch up the Plain Vanilla Scheme

In our experience discussing these issues, there is a very strong and proper urge to consider the possibility that we could—perhaps in some cursory way as suggested by Berger and Wolpert—paper over the difficult aspects of the DSGE context and somehow follow some scheme that has a reasonable pragmatic interpretation as being consistent with the plain vanilla scheme.

Obviously, these suggestions involve methods to eliminate *ad hoc* aspects of the prior and to accommodate in some way the most gross deficiencies of the model. For example, Del Negro and Schorfheide (2008) what appears to be a very useful approach to training samples as a method to reducing the sort of problems with the conventional DSGE prior that they emphasize and that we have illustrated above. Geweke has recently provided a brilliant monograph on working with complete and incomplete econometric models, which involves methods for comparing explicit models based on particular features and setting aside certain types of deficiency.

All these strike us as very good ideas that should find widespread uses. We argue that that are complementary to our suggestions, however. Such steps cannot overturn the fact that for the foreseeable future, our best complete general equilibrium model will be materially deficient, in an ongoing state of refinement on a single dataset, and simultaneously in use in policymaking. We argue that under these conditions, any coherent class of analysis must allow for the examination of the flaws of the current model+*posterior* that is informing the policymaking process.

5 Conclusions

The modelling of causal channels in a large, dynamic system, with forward looking behavior is an incredibly daunting task. It is made even more challenging by the fact that we have a single dataset on which to both develop and test our theories. That dataset is small in relevant senses.

Perhaps it is not surprising in such a challenging context that there is little consensus in the profession on key substantive economic topics or on modelling methodology. The unprecedented Congressional hearing on DSGE modelling underscores both the unfortunate state of affairs and the fact that improving on this state of affairs is in the interest of everyone subject to the policy decisions informed by these models.

We argue for taking as a starting point that these models will actively be used in policy analysis while remaining in an ongoing state of material refinement. Seen in this light, posterior predictive analysis, we argue, can be a very useful tool for highlighting strengths and weaknesses pertinent to policy. We particularly argue for the focus on what we call structural features as a way to assess causal channels in the model. This type of analytics has received little prior use in DSGE modelling. The overarching idea is that the the type of analysis we advocate may serve to inform policymakers of limitations of the current models (which can then be judgementally allowed for) and to direct resources of the model refinement efforts to areas most relevant to policymaking.

References

Attanasio, O., Weber, G., 1993. Consumption growth, the interest rate and aggregation. The Review of Economic Studies 60 (3), 631–649.

Bayarri, M., Berger, J., 1999. Comment on Bayarri and Berger. Bayesian Statistics 6, 53–67.

Berger, J., Wolpert, R., 1984. The likelihood principle. Institute of Mathematical Statistics.

Bernanke, B., Boivin, J., 2003. Monetary policy in a data-rich environment. Journal of Monetary Economics 50 (3), 525–546.

Bernanke, B., Boivin, J., Eliasz, P., 2005a. Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. Quarterly Journal of Economics 120 (1), 387–422.

Bernanke, B., Reinhart, V., Sack, B., 2005b. Monetary policy alternatives at the zero bound: An empirical assessment. Brookings Papers on Economic Activity 2004 (2), 1–100.

Bernardo, J., 1999. Quantifying surprise in the data and model verification. Bayesian Statistics 6, 72–73.

Box, G., 1980. Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society. Series A (General) 143 (4), 383–430.

Camerer, C., 1995. Individual decision making. The handbook of experimental economics 3, 587–704.

Carroll, C., Slacalek, J., Sommer, M., forthcoming. International evidence on sticky consumption growth. The Review of Economics and Statistics.

Chari, V., July 2010. Testimony before the U.S. House of Representatives. House Committee on Science and Technology, Subcommittee on Investigations and Oversight.

Christiano, L., Eichenbaum, M., Evans, C., 2005. Nominal rigidities and the dynamic effects of a monetary policy shock. Journal of Political Economy 113 (1), 1–45.

Del Negro, M., Schorfheide, F., 2008. Forming priors for DSGE models (and how it affects the assessment of nominal rigidities). Journal of Monetary Economics 55 (7), 1191–1208.

Del Negro, M., Schorfheide, F., Smets, F., Wouters, R., 2007. On the fit of New Keynesian models. Journal of Business & Economic Statistics 25 (2), 123–143.

Dynan, K., 2000. Habit formation in consumer preferences: Evidence from panel data. American Economic Review, 391–406.

Faust, J., 2009. The New Macro Models: Washing Our Hands and Watching for Icebergs. Economic Review, 45–68.

Faust, J., Gupta, A., 2010b. Are all recessions black swans? DSGE models and the post-war U.S. business cycle. in progress, Johns Hopkins University.

Faust, J., Wright, J., 2009. Comparing Greenbook and reduced form forecasts using a large realtime dataset. Journal of Business and Economic Statistics 27 (4), 468–479.

Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2005. The generalized dynamic factor model. Journal of the American Statistical Association 100 (471), 830–840.

Gelman, A., Meng, X., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6, 733–759.

Geweke, J., 2005. Contemporary Bayesian econometrics and statistics. John Wiley.

Geweke, J., 2007. Bayesian model comparison and validation. American Economic Review 97 (2), 60–64.

Geweke, J., 2010. Complete and incomplete econometric models. monograph, Princeton University Press.

Giannone, D., Reichlin, L., Sala, L., 2004. Monetary policy in real time. NBER Macroeconomics Annual 19, 161–200.

Gilchrist, S., Yankov, V., Zakrajsek, E., 2009. Credit market shocks and economic fluctuations: Evidence from corporate bond and stock markets. Journal of Monetary Economics 56 (4), 471–493.

Gupta, A., 2010. A forecasting metric for evaluating DSGE models for policy analysis. in progress, Johns Hopkins University.

Hansen, B., 2005. Challenges for econometric model selection. Econometric Theory 21 (01), 60–68.

Harvey, A., 1991. Forecasting, structural time series models and the Kalman filter. Cambridge Univ Pr.

Hill, B., 1996. Comment on Gelman, et al. Statistica Sinica 6, 767–773.

Kydland, F., Prescott, E., 1982. Time to build and aggregate fluctuations. Econometrica 50 (6), 1345–1370.

Lancaster, T., 2004. An introduction to modern Bayesian econometrics. Wiley-Blackwell.

Mehra, R., Prescott, E., 1985. The equity premium: A puzzle. Journal of monetary Economics 15 (2), 145–161.

Sims, C., 1980. Macroeconomics and reality. Econometrica: Journal of the Econometric Society 48 (1), 1–48.

Smets, F., Wouters, R., 2003. An estimated dynamic stochastic general equilibrium model of the Euro area. Journal of the European Economic Association 1 (5), 1123–1175.

Smets, F., Wouters, R., 2007. Shocks and frictions in US business cycles: A Bayesian DSGE approach. The American Economic Review 97 (3), 586–606.

Solow, R., July 2010. Building a science of economics for the real world. House Committee on Science and Technology, Subcommittee on Investigations and Oversight.

Stock, J., Watson, M., 1999. Forecasting inflation. Journal of Monetary Economics 44 (2), 293–335.

Stock, J., Watson, M., 2002. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97 (460), 1167–1179.

Stock, J., Watson, M., 2003. Forecasting output and inflation: the role of asset prices. Journal of Economic Literature 41 (3), 788–829.

Stock, J., Watson, M., 2005. An empirical comparison of methods for forecasting using many predictors. Manuscript, Princeton University.

Tiao, G., Xu, D., 1993. Robustness of maximum likelihood estimates for multi-step predictions: the exponential smoothing case. Biometrika 80 (3), 623–641.

Wilcox, D., 1992. The construction of US consumption data: Some facts and their implications for empirical work. The American economic review 82 (4), 922–941.

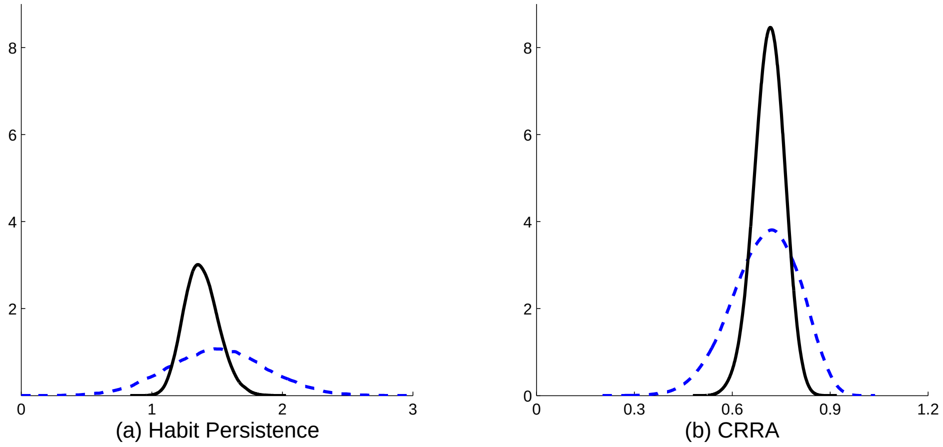
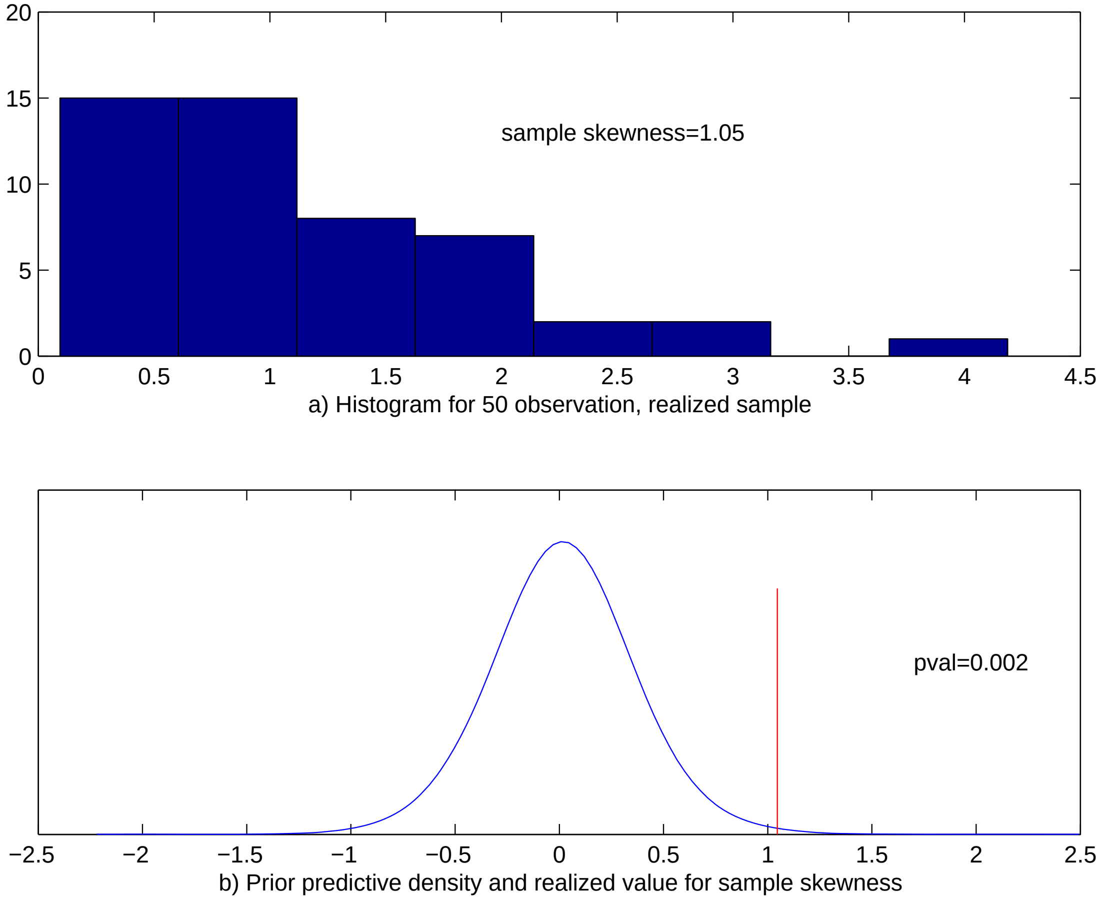


Figure 1: Prior and posterior densities for habit persistence and CRRA parameters. Blue dashed line is the prior; black line is the posterior. The parameters are described more fully in the text.



a) Histogram for 50 observation, realized sample

b) Prior predictive density and realized value for sample skewness

Figure 2: Example of prior predictive analysis. Panel a) Histogram of 50 sample observations. Panel b) The prior predictive density for sample skewness and realized value for sample. The prior is that the sample points are iid $N(\mu, 1)$ and the prior for μ is uniform on $[0, 1]$. The sample skewness of 1.05 is shown in red on the right hand panel. The stated p-value is the share of the mass of the prior predictive density exceeding the realized sample skewness.

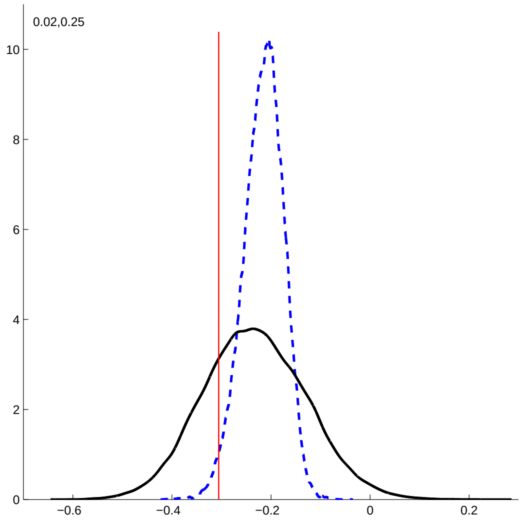
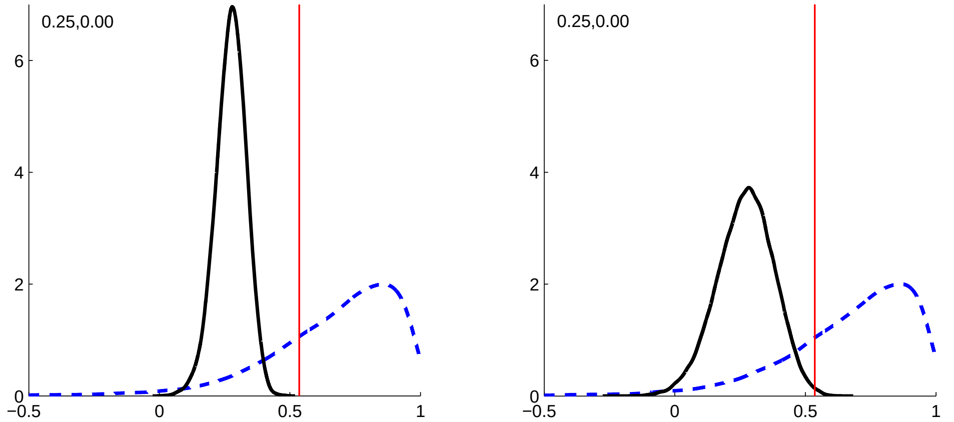


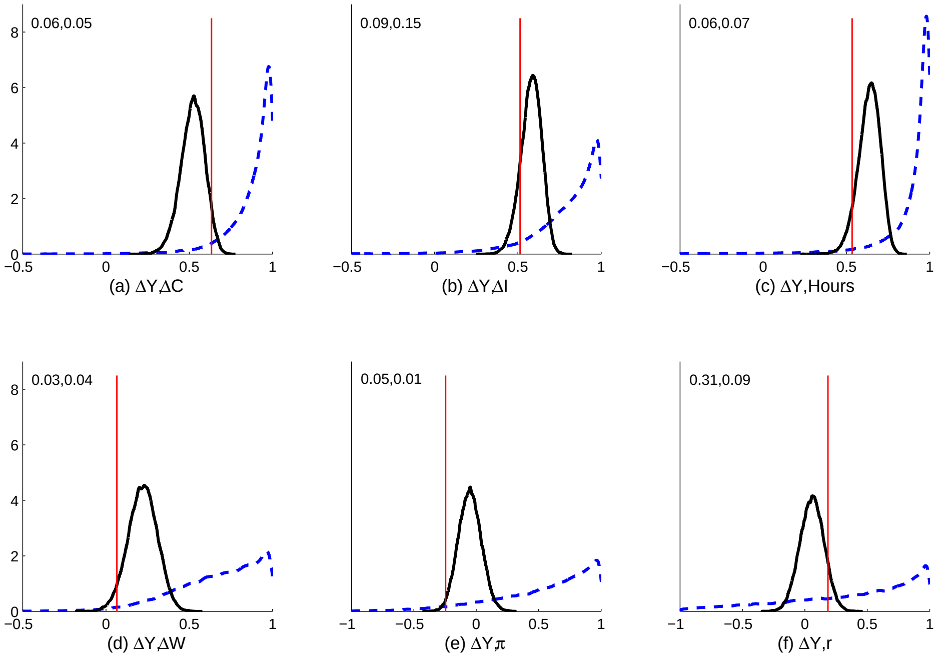
Figure 3: Posterior density for population correlation of output and inflation (blue dashed) and posterior predictive density for the sample correlation (black solid). Red line is the realized value on the sample. The numbers in the upper left give the proportion of mass under the posterior and posterior predictive density, respectively, that is to the left of the realized value.



(a) Prior and Posterior for Population Correlation

(b) Prior Predictive and Posterior Predictive for Sample Correlation

Figure 4: Correlation of consumption and investment growth. Panel a) Prior (blue dashed) and posterior (black solid) for population correlation. Panel b) Prior predictive (blue dashed) and posterior predictive (black solid) for sample correlation. The numbers in the upper left give the share of points in the smaller tail relative to the red line for the two densities on the panel, with value for prior before posterior.



(a) ΔY,ΔC

(b) ΔY,ΔI

(c) ΔY,Hours

(d) ΔY,ΔW

(e) ΔY,π

(f) ΔY,r

Figure 5: Prior predictive (dashed blue) and posterior predictive (solid black) densities for one-step forecast error correlations. Each panel is for a sample correlation of the output growth error ΔY and the error for one of the other variables in the model: ΔC, consumption growth; ΔI, investment growth; hours, Δw, wage growth; π, inflation; r, interest rate. The errors come from a VAR(1) estimated on the sample. Red line is for the VAR(1) value on the realized sample. The numbers in the upper left give the share of points in the smaller tail relative to the red line for the two densities on the panel, with value for prior before posterior.

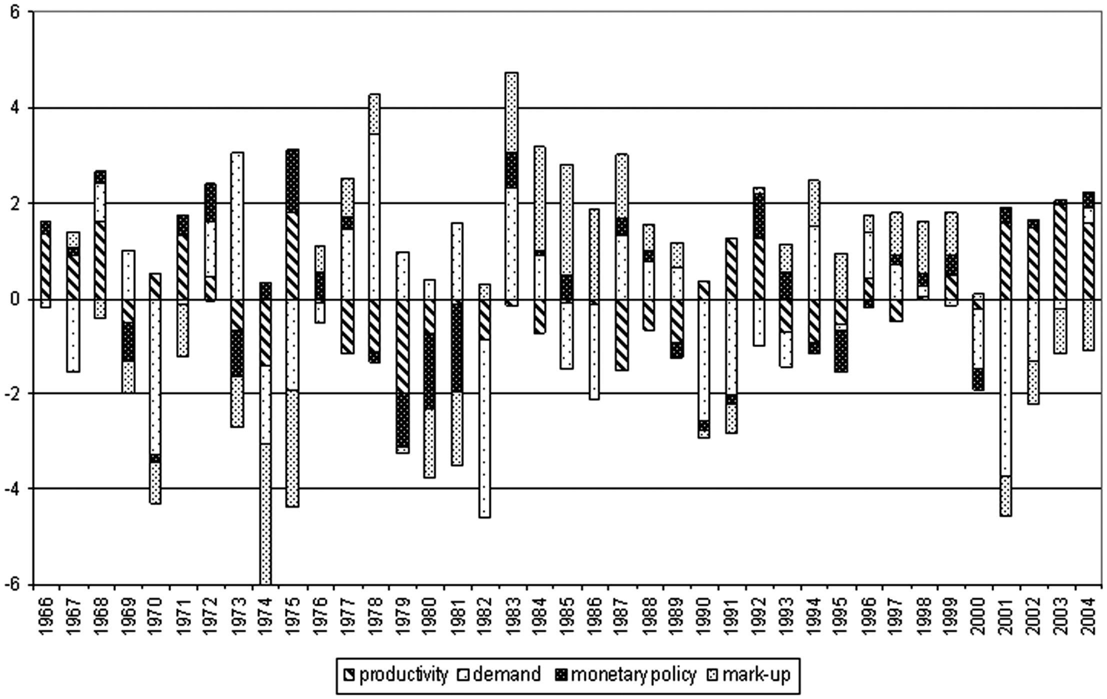
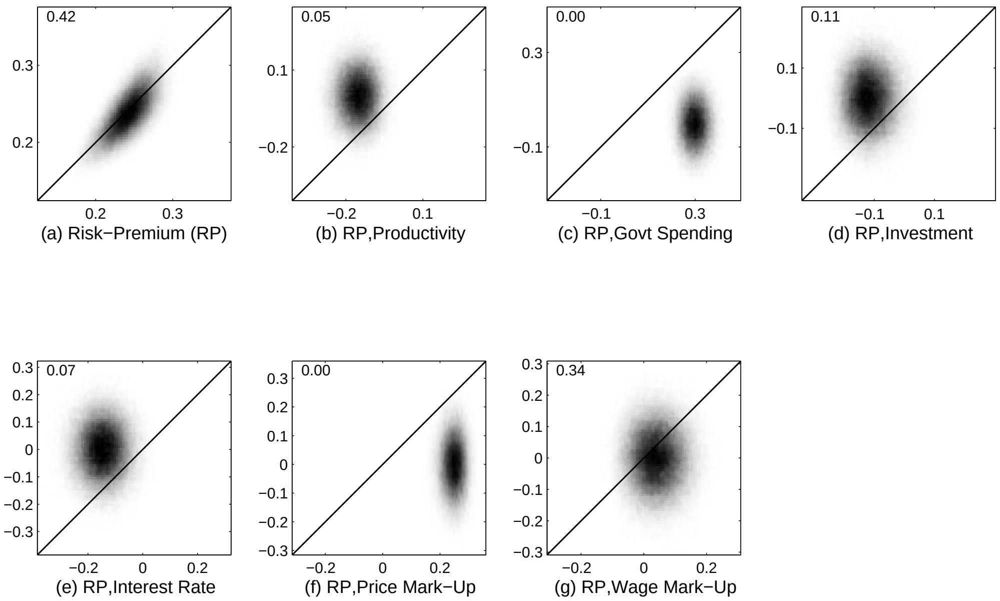


Figure 6: Historical decomposition of output growth in terms of the structural shocks. The 7 shocks have been averaged over calendar years and summed across broad categories. The 'demand shocks' include the risk premium, investment-specific technology, and exogenous spending shocks; the 'mark-up shocks' include the price and wage mark-up shocks. Source: Smets-Wouters (2007).



(a) Risk−Premium (RP)

(b) RP,Productivity

(c) RP,Govt Spending

(d) RP,Investment

(e) RP,Interest Rate

(f) RP,Price Mark−Up

(g) RP,Wage Mark−Up

Figure 7: Structural feature scatter plots for the smoothed risk-premium (rp) shock. Panel a) is for the standard deviation of the rp shock; the remaining panels are for the sample correlation of the rp shock with other structural shocks in the model: productivity, investment productivity, government spending, monetary policy, price mark-up, and wage mark-up. Horizontal axis plots the posterior density for the realized sample; vertical axis plots the posterior predictive values. The number in the upper left gives the smaller share of points on either side of the 45 degree line.

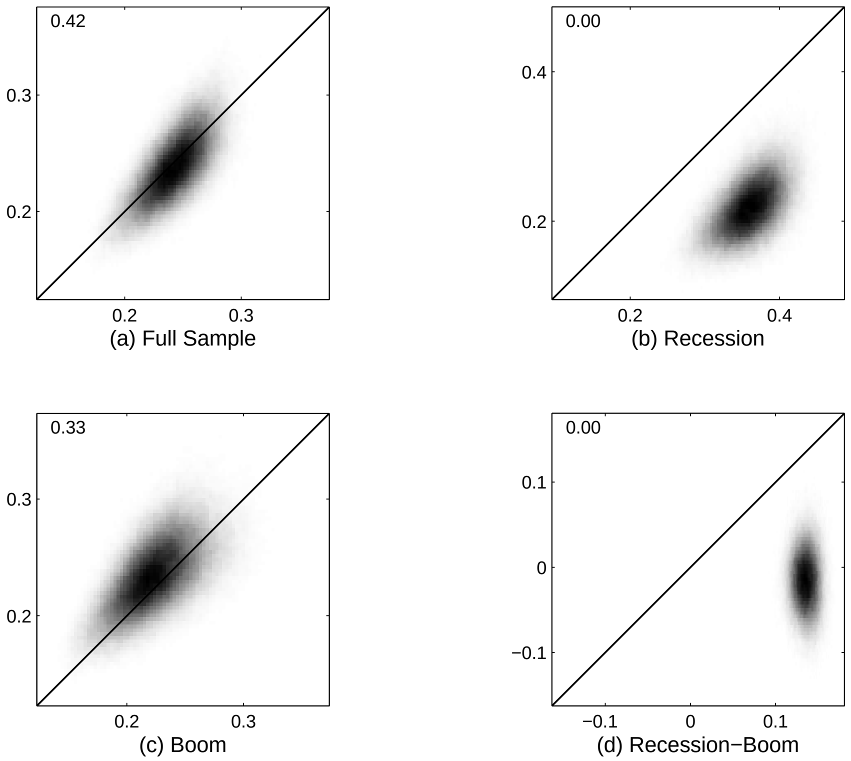


Figure 8: Structural feature scatter plot for the sample standard deviation of the smoothed risk-premium shock in recessions and expansions. Panel (a) is for the full sample, (b) recessions, (c) expansions. In panel (d), the feature is the difference in the standard deviation for recessions and expansions. Horizontal axis plots the posterior values for the realized sample; vertical axis plots the posterior predictive values. The number in the upper left gives the smaller share of points on either side of the 45 degree line.