On the Determination of General Scientific Models with Application to Asset Pricing [1]

A. Ronald Gallant
Fuqua School of Business
Duke University
Durham NC 27708-0120 USA

Robert E. McCulloch
Graduate School of Business
University of Texas
Austin TX 78712-1175 USA

Abstract

We consider a consumption based asset pricing model that uses habit persistence to overcome the known statistical inadequacies of the classical consumption based asset pricing model. We find that the habit model fits reasonably well and agrees with results reported in the literature if conditional heteroskedasticity is suppressed but that it does not fit nor do results agree if conditional heteroskedasticity, well known to be present in financial market data, is allowed to manifest itself. We also find that it is the preference parameters of the model that are most affected by the presence or absence of conditional heteroskedasticity, especially the risk aversion parameter. The habit model exhibits four characteristics that are often present in models developed from scientific considerations: (1) a likelihood is not available; (2) prior information is available; (3) a portion of the prior information is expressed in terms of functionals of the model that cannot be converted into an analytic prior on model parameters; (4) the model can be simulated. The underpinning of our approach is that, in addition, (5) a parametric statistical model for the data, determined without reference to the scientific model, is known. In general one can expect to be able to determine a model that satisfies (5) because very richly parameterized statistical models are easily accommodated. We develop a computationally intensive, generally applicable, Bayesian strategy for estimation and inference for scientific models that meet this description together with methods for assessing model adequacy. An important adjunct to the method is that a map from the parameters of the scientific model to functionals of the scientific and statistical models becomes available. This map is a powerful tool for understanding the properties of the scientific model.

Keywords: Scientific models, simulation, Bayes, MCMC, estimation, inference, asset pricing.

1 Introduction

This article is motivated by an asset pricing application, namely the habit persistence asset pricing model, that has characteristics in common with all modern general equilibrium models. And this article contributes to that empirical literature. But many models derived from the principles of the physical, biological, social, or engineering sciences exhibit the same characteristics that make statistical analysis of the habit model challenging. These characteristics are (1) a likelihood is not available; (2) prior information is available; (3) a portion of the prior information is expressed in terms of functionals of the model that cannot be converted into an analytic prior on model parameters; (4) the model can be simulated. We outline a general approach for the analysis of such models and then apply it to the habit model.

In some instances other methods are available. For example, if the only cause of difficulties is a small number of latent variables, then a data augmentation approach will likely be applicable and be less computationally intensive than the methods proposed here. We are not concerned with such models. Our concern is models where there is little else available. Often the serious use of prior information is essential because the available data are sparce. Our proposals incorporate prior information and are especially helpful when some of this prior information can only be expressed in terms of functionals of the model. In our application, data are sparse, there seems little else available that can incorporate prior information, and some prior information must be expressed in terms of funcationals. The methods proposed here also generate ancillary information that can help to interpret the scientific model in terms of its statistical properties and to cast model inadequacies into sharp relief.

Our approach depends on an assumption that (5) an adequate statistical model for the data is available. Because the statistical model is only estimated in large simulated data sets, richly parameterized models may be used so that (5) can usually be satisfied. Briefly, our proposal is as follows. Given (5), we can construct a map from the parameters of the scientific model to those of the statistical model such that a point in the parameter space of the scientific model and its image under the map both correspond to the same data generating process. Typically the parameters of the statistical model will live in a higher dimensional space than that of the scientific model. In Section 3 we obtain a Bayesian inference for the scientific model using the map to compute the likelihood. In Section 4, we work in terms

of the statistical model and its parameters, but use the map as a key element in our prior construction. This construction allows us to assess the adequacy of the scientific model. The methodology developed here allows the prior information to be expressed either directly on the parameters of the scientific model or on functionals of the scientific model that can be evaluated via simulation.

The idea that we intend to convey by the term scientific model is that the model is discipline-based and statistically intractable. The idea that we intend to convey by the term statistical model is that it has been obtained by data-analytic considerations with the specific intent of being statistically tractable.

The discovery of the mapping from the parameters of the scientific model to those of the statistical model, which is an intermediate step of the methods proposed here, is often itself of interest. For instance, the statistical model must, perforce, be expressed entirely in terms of observables whereas scientific models often contain unobservables. Having a mapping from the subset of the parameters that control the unobservable features of the scientific model to the parameters of a statistical model consisting entirely of observables can be helpful in understanding the observable consequences of changes in a model's unobservable internal structure. The utility of this approach can be extended by using the same methods to find the map from the parameters of the scientific model to functionals of both the scientific and statistical models.

A Bayesian approach suggests itself for problems that exhibit the five characteristics just listed because the methodology gracefully accepts prior information into the analysis. Nonetheless, comparison is of interest and in the application we compare to the results of Bansal, Gallant, and Tauchen (2007), referred to as BGT hereafter, who use a synthesis of frequentist methods proposed by Smith (1993), Gourieroux, Monfort, and Renault (1993), and Gallant and Tauchen (1996). This approach is logically distinct from ours but does make use of an auxiliary statistical model as an adjunct to estimation and inference as we do here. As will be seen, sparse data force a simplification on BGT estimates that inhibits discovery of our findings. See also Del Negro and Schorfheide (2004), Dejong, Ingram, Whiteman (1996, 2000) for closely related ideas under linearity assumptions and the references therein. Our implementation relies on modern object oriented programming methods, modern data structures, and a discretization at a critical point in the computations. Bringing these elements to bear on the problem seems to be both novel to this work and essential to success.

We believe that our proposals for model assessment are new.

2 Scientific and Statistical Models

We shall use the notational conventions of time series analysis because most models of the sort considered here are dynamic. This is in no way essential because the results apply equally well to other data with a few obvious changes to notation.

Let the transition density of the scientific model be denoted as $p(\boldsymbol{y}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{x}_{t-1} = (\boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-L})$ if Markovian and $\boldsymbol{x}_{t-1} = (\boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_1)$ if not. We assume that there is no explicit mathematical representation of $p(\cdot|\cdot, \boldsymbol{\theta})$. All that we can do is simulate data from $p(\cdot|\cdot, \boldsymbol{\theta})$ for given $\boldsymbol{\theta}$. If the model produces ergodic output, then a single long simulation for each setting of $\boldsymbol{\theta}$ suffices for our purposes. If not, then many independently simulated replicates of the data are used.

Since here is no closed form expression for $p(\cdot|\cdot, \boldsymbol{\theta})$, there is no direct way to compute the likelihood. Our approach is to find a parametric family of distributions that is capable of representing the process $\{\boldsymbol{y}_t\}$:

ASSUMPTION 1 We assume that there is a transition density $f(\boldsymbol{y}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\eta})$, $\boldsymbol{\eta} \in \boldsymbol{H}$, and that there is a one-to-one map $g : \boldsymbol{\theta} \mapsto \boldsymbol{\eta}$ such that

$$p(\boldsymbol{y}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta}) = f(\boldsymbol{y}_t|\boldsymbol{x}_{t-1}, \boldsymbol{g}(\boldsymbol{\theta})) \qquad \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{1}$$

and that the form of $f(\cdot|\cdot, \boldsymbol{\eta})$ is known.

When we need a likelihood based on the unknown $p(\cdot|\cdot, \boldsymbol{\theta})$, we substitute $f(\cdot|\cdot, \boldsymbol{g}(\boldsymbol{\theta}))$. The model $f(\cdot|\cdot, \boldsymbol{\eta})$ is a statistical description of the observed data that we call the statistical model. Often this model will be known from the literature. In other cases it must be determined as part of the analysis. As richly parameterized models are permitted, success in finding an acceptable statistical model can be anticipated. It is to be emphasized that we only use the statistical model to fit large simulations from the scientific model (Section 3) or when augmented by a strong prior dictated by the scientific model (Section 4) so that the fact that the data may be too sparse to support it is not a consideration.

When Assumption 1 is satisfied the likelihood is exactly that implied by the scientific model. When Assumption 1 is violated the likelihood is different from that implied by the

scientific model. To use Poirier's (1988) terminology, when Assumption 1 holds one is looking at the world through the window implied by the scientific model. When Assumption 1 is violated one is looking at the world through a different window.

One might deliberately choose to violate Assumption 1. For example, if satisfaction of (1) leads to a statistical model $f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ that differs markedly from the distribution of the data, one might deliberately opt for a simplification of $f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ that does not exhibit these characteristics. The issues that arise in this connection are discussed in general in Subsection 3.4 and specifically for our application in Subsections 5.4 and 5.7.

The scientific model is built using subject matter knowledge. Thus, we expect that prior information is available. This prior information may be expressible either in terms of elements of $\boldsymbol{\theta}$ or in terms of characteristics $\boldsymbol{\psi}$ of the process. For example, in our application an element of $\boldsymbol{\psi}$ is the unconditional moment of a variable for which no data that corresponds to its meaning within the model exists. In general, $\boldsymbol{\psi}$ can be regarded as a point in the range of a vector of functionals

$$\boldsymbol{\Psi} : p(\cdot|\cdot, \boldsymbol{\theta}) \mapsto \boldsymbol{\psi} \tag{2}$$

that is computable from a simulation. Thus, $\boldsymbol{\psi}$ is a function of $\boldsymbol{\theta}$ through the composition $\boldsymbol{\theta} \mapsto p(\cdot|\cdot, \boldsymbol{\theta}) \mapsto \boldsymbol{\psi}$. We capture both of these types of information in our prior $\pi(\boldsymbol{\theta})$ through the construction

$$\pi(\boldsymbol{\theta}) \propto h(\boldsymbol{\theta}, \boldsymbol{\Psi}(\boldsymbol{\theta})). \tag{3}$$

Note that since we will be using the Metropolis-Hastings algorithm to compute the posterior, we only need a function proportional to the prior. We shall also discretize $\boldsymbol{\theta}$ on a finite grid so that any positive h will be integrable.

<center>Figure 1 about here</center>

We may not want to impose the belief that the scientific model holds exactly. We can capture this idea by recasting the problem so that $\boldsymbol{\eta}$ of the statistical model is viewed as the parameter of interest and constructing a prior that expresses a preference for $\boldsymbol{\eta}$ that are close to the manifold

$$\boldsymbol{\mathcal{M}} = \left\{ \boldsymbol{\eta} \in \boldsymbol{H} \ : \ \boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta}), \, \boldsymbol{\theta} \in \boldsymbol{\Theta} \right\}, \tag{4}$$

where $\boldsymbol{\Theta}$ is the parameter space of the scientific model and $\boldsymbol{H}$ is the parameter space of the statistical model. Our prior construction uses a single parameter that we call κ to control prior beliefs about how close $\boldsymbol{\eta}$ should be to the manifold (Section 4). The smaller κ is,

<center>4</center>

the more prior weight is placed on $\boldsymbol{\eta}$ close to the manifold. We assess the scientific model by seeing if the marginal posterior distributions of interpretable features of the statistical model are sensitive to the choice of κ. If changing our prior so as to support $\boldsymbol{\eta}$ farther from the manifold results in location shifts of the posteriors that are appreciable from a practical point of view, then we conclude that the evidence in the likelihood is against the restriction corresponding to the scientific model.

We illustrate the ideas in Figure 1. The scientific model is $p(y|\theta) = n(y; \theta, \theta^2)$, and the statistical model is $f(y|\boldsymbol{\eta}) = n(y; \eta_1, \eta_2)$, where $n(y; \mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2. The mapping of the parameter θ of the scientific model to the parameters (η_1, η_2) of the statistical model is $\boldsymbol{g} : \theta \mapsto (\theta, \theta^2)$. The panels of Figure 1 are as follows: (i) the curve depicts the manifold (4), (ii) the dotted contours are those of the likelihood of the statistical model, (iii) the shading depicts the prior on $\boldsymbol{\eta}$ corresponding to a choice of κ, (iv) the dots are draws from the posterior for the parameter θ of the scientific model mapped into $\boldsymbol{\mathcal{M}} \in \boldsymbol{H}$ using the map $\boldsymbol{\eta} = \boldsymbol{g}(\theta)$ and then jittered (without the jittering, all the draws would be on the manifold), and (v) the solid contours represent the posterior of $\boldsymbol{\eta}$ given the prior (iii).

The likelihood in the two left panels was obtained by simulating 50 observations from the scientific model with $\theta = 2$, or equivalently, $(\eta_1 = 2, \eta_2 = 4)$. The likelihood in the two right panels was obtained by simulating 50 observations from the statistical model with $\boldsymbol{\eta} = (2.8, 4)$. The prior off the manifold corresponds to a small value of κ in the top two panels, and a larger value in the bottom two. Our method for assessment of the scientific model is based on the observation that when the data support the scientific model, increasing κ may cause the posterior to become more spread out, but it will not dramatically shift location (panels (1,1) and (2,1)). Conversely, when the data does not support the scientific model, increasing κ will result in a shift of the posterior (panels (1,2) and (2,2)).

Of course, in high dimensional problems, we cannot simply look at the contours of the $\boldsymbol{\eta}$ posterior. Instead we must examine low dimensional marginals of interest. The densities depicted with the solid curves in the (1,2) and (2,2) panels of Figure 5 show the effect of increasing κ on a marginal of interest in our asset-pricing application. We view the shift in the distribution as evidence against the model.

At this point, the main conceptual ideas that we shall propose have been set forth. The devil is in the details, to which we now proceed. The reader who would rather see our

substantive results first can skip to Section 5.

3 Bayesian Estimation of Scientific Models

We have two cases to consider. In the first case, we assume the scientific model is true and seek inference for $\boldsymbol{\theta}$. Here the statistical model is just a tool for computing the likelihood. In the second case we work in the context of the statistical model and $\boldsymbol{\eta}$ is the parameter. In this case, the scientific model is a source of prior information. We will consider the first case in this section and the second in Section 4. Note that we are never working on the product space $\boldsymbol{\Theta} \times \boldsymbol{H}$ as is the case in work closely related to ours (Del Negro and Schorfheide 2004).

3.1 A Metropolis Algorithm for θ

We use the Metropolis algorithm to compute the posterior distribution of $\boldsymbol{\theta}$. The Metropolis algorithm is an iterative scheme generating a sequence of $\boldsymbol{\theta}$ values according to a Markov chain whose stationary distribution is the posterior. To implement it, we must require a likelihood, a prior, and transition density in $\boldsymbol{\theta}$ called the proposal density.

Let $\mathcal{L}(\boldsymbol{\theta})$ denote the likelihood assuming that (1) holds. To compute it we use

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{t=1}^{n} f(\boldsymbol{y}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{g}(\boldsymbol{\theta})),$$

where $\{\boldsymbol{y}_t, \boldsymbol{x}_{t-1}\}$ denotes the observed data and n the sample size. Let $\pi(\boldsymbol{\theta})$ denote the prior distribution on $\boldsymbol{\theta}$. As discussed in Section 2, in order to compute this prior $\pi(\boldsymbol{\theta})$ we may need the value $\boldsymbol{\psi}$ taken on by the functionals $\boldsymbol{\Psi}$ given by (2). Let q denote our proposal density. For a given $\boldsymbol{\theta}^o$, $q(\boldsymbol{\theta}^o, \boldsymbol{\theta}^*)$ defines the distribution of a potential new value $\boldsymbol{\theta}^*$.

Given a current $\boldsymbol{\theta}^o$ and the corresponding $\boldsymbol{\eta}^o = \boldsymbol{g}(\boldsymbol{\theta}^o)$, we obtain the next pair $(\boldsymbol{\theta}', \boldsymbol{\eta}')$ as follows:

1. Draw $\boldsymbol{\theta}^*$ according to $q(\boldsymbol{\theta}^o, \boldsymbol{\theta}^*)$.

2. Draw $\{\hat{\boldsymbol{y}}_t, \hat{\boldsymbol{x}}_{t-1}\}_{t=1}^{N}$ according to $p(\boldsymbol{y}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta}^*)$.

3. Compute $\boldsymbol{\eta}^* = \boldsymbol{g}(\boldsymbol{\theta}^*)$ and $\boldsymbol{\psi}^*$ from the simulation $\{\hat{\boldsymbol{y}}_t, \hat{\boldsymbol{x}}_{t-1}\}_{t=1}^{N}$.

4. Let $\alpha = \min\left(1, \frac{\mathcal{L}(\boldsymbol{\theta}^*)\,\pi(\boldsymbol{\theta}^*)\,q(\boldsymbol{\theta}^*, \boldsymbol{\theta}^o)}{\mathcal{L}(\boldsymbol{\theta}^o)\,\pi(\boldsymbol{\theta}^o)\,q(\boldsymbol{\theta}^o, \boldsymbol{\theta}^*)}\right)$.

5. With probability α, $(\boldsymbol{\theta}', \boldsymbol{\eta}') = (\boldsymbol{\theta}^*, \boldsymbol{\eta}^*)$, otherwise $(\boldsymbol{\theta}', \boldsymbol{\eta}') = (\boldsymbol{\theta}^o, \boldsymbol{\eta}^o)$.

Steps 1, 4, and 5 are just the standard Metropolis algorithm. Steps 2 and 3 are essential features of our approach. If the proposed $\boldsymbol{\theta}$ in Step 1 violates a support condition that can be checked without running Step 2, one skips Step 2 because α in Step 4 will be zero.

3.2   Choice of $\boldsymbol{\theta}$ Proposal

Step 1 of the Metropolis algorithm in Subsection 3.1 requires a proposal density q for $\boldsymbol{\theta}$. In choosing q we need to take into consideration that to compute the likelihood at a proposed $\boldsymbol{\theta}$ the scientific model must be simulated. For a sophisticated scientific model, this simulation may involve significant computation. A nonlinear optimizer, a nonlinear equation solver, or some other routine that needs starting values might be called in the course of these computations. This motivates us to consider proposing small changes in $\boldsymbol{\theta}$ so that computational results from the old $\boldsymbol{\theta}$ may be used in doing the computations for the proposed $\boldsymbol{\theta}$. In particular, if $\boldsymbol{\theta}$ is not changed too much, results from the previous computation can be used as starting values for the new one. But this strategy creates more dependence in the Markov chain.

We start by discretizing $\boldsymbol{\theta}$ because, as seen later, discretization permits significant improvements in computational efficiency. For the ith component of $\boldsymbol{\theta}$ we choose $a_i < b_i$, and s_i. We then let θ_i take on the values $a_i + js_i$ where j ranges from 1 to g_i equal to the integer part of $(b_i - a_i)/s_i$. Thus, θ_i takes values between a_i and b_i on a grid of mesh s_i. To propose a new $\boldsymbol{\theta}$ we first randomly choose a component to change, with each component having the same chance of being chosen. If the ith component is chosen, there is some j such that the current $\theta_i = a_i + js_i$. We choose a set of distributions $q_i(j, k)$ on $\{1, 2, \ldots, g_i\}$ where i is the $\boldsymbol{\theta}$ component, j is the current grid position of that component, and k denotes the random new grid position to be drawn. We draw $k \sim q_i(j, \cdot)$ and let $\boldsymbol{\theta}^*$ be obtained from $\boldsymbol{\theta}$ by changing the ith component from $a_i + js_i$ to $a_i + ks_i$. For $q_i(j, k)$ use

$$
q_i(j, k) \propto \begin{cases} \exp(-\frac{1}{2\sigma_i^2}(k - j)^2) & k \neq j \\ 0 & \text{else} \end{cases}.
$$

The choice of σ_i determines the number of s_i that we tend to move. We assign 0 probability to proposing that we stay put as there is no point in proposing that we go to where we are.

The choice of a_i and b_i is not critical; a_i and b_i can be set so that the intervals (a_i, b_i) cover the support of the posterior by a wide margin without noticeably degrading the performance

of the Metropolis algorithm. The choice of s_i is crucial. We will move away from the starting value in integer multiples of s_i. The combination of the choice of s_i and σ_i determines the size of the changes that q proposes. The choice of s_i determines the accuracy of our inference. When we choose s_i we are saying that, as a practical matter, we only need to know θ_i in terms of s_i units. Two $\boldsymbol{\theta}$'s that differ in component i by less than s_i are virtually the same as a practical matter. Since computation is expensive, we should not waste resources by determining $\boldsymbol{\theta}$ on a finer scale than we actually care about.

3.3   Computing the Map

In Step 3 of the Metropolis algorithm in Subsection 3.1 we need to uncover the map $\boldsymbol{g} : \boldsymbol{\theta} \mapsto \boldsymbol{\eta}$ that satisfies (1) from a simulation $\{\hat{\boldsymbol{y}}_t, \hat{\boldsymbol{x}}_{t-1}\}_{t=1}^{N}$ of $p(\boldsymbol{y}_t | \boldsymbol{x}_{t-1}, \boldsymbol{\theta})$. At an intuitive level what we propose is simple: We choose N so large that the simulated data $\{\hat{\boldsymbol{y}}_t, \hat{\boldsymbol{x}}_{t-1}\}_{t=1}^{N}$ gives us complete information about its distribution for given $\boldsymbol{\theta}$. We find the corresponding $\boldsymbol{\eta}$ by maximizing the likelihood of the simulated data under the statistical model $f(\cdot | \cdot, \boldsymbol{\eta})$. That is, we find the $\boldsymbol{\eta}$ which gives the same kind of data under $f(\boldsymbol{y}_t | \boldsymbol{x}_{t-1}, \boldsymbol{\eta})$ as did $\boldsymbol{\theta}$ under $p(\boldsymbol{y}_t | \boldsymbol{x}_{t-1}, \boldsymbol{\theta})$. At a formal level, we are finding the $\boldsymbol{\eta}$ that puts the Kullback-Liebler divergence $d(f, p) = \iint [\log p(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\theta}) - \log f(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\eta})] \, p(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\theta}) \, d\boldsymbol{y} \, p(\boldsymbol{x} | \boldsymbol{\theta}) \, d\boldsymbol{x}$ to zero by minimizing $d(f, p)$ with respect to $\boldsymbol{\eta}$ and are noting that $\iint \log p(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\theta}) \, p(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\theta}) \, d\boldsymbol{y} \, p(\boldsymbol{x} | \boldsymbol{\theta}) \, d\boldsymbol{x}$ does not have to be computed to solve this minimization problem. We approximate the integral that does have to be computed in the usual way: $\iint \log f(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\eta}) \, p(\boldsymbol{y} | \boldsymbol{x}, \boldsymbol{\theta}) \, d\boldsymbol{y} \, p(\boldsymbol{x} | \boldsymbol{\theta}) \, d\boldsymbol{x} \approx \frac{1}{N} \sum_{t=1}^{N} \log f(\hat{\boldsymbol{y}}_t | \hat{\boldsymbol{x}}_{t-1}, \boldsymbol{\eta})$. (Or by $\frac{1}{R} \sum_{r=1}^{R} \frac{1}{n} \sum_{t=1}^{n} \log f(\hat{\boldsymbol{y}}_{t,r} | \hat{\boldsymbol{x}}_{t-1,r}, \boldsymbol{\eta})$ if not ergodic. We assume ergodicity hereafter; if not, the requisite modifications are obvious.) Thus, upon dropping the division by N, the map is computed as

$$\boldsymbol{g} : \boldsymbol{\theta} \mapsto \operatorname*{argmax}_{\boldsymbol{\eta}} \sum_{t=1}^{N} \log f(\hat{\boldsymbol{y}}_t \,|\, \hat{\boldsymbol{x}}_{t-1}, \boldsymbol{\eta}).$$

Since the $f(\cdot | \cdot, \boldsymbol{\eta})$ family is generally chosen to be flexible and high dimensional, this likelihood can be complicated. However, the simulated data set is large and $\boldsymbol{\eta}^o$ should be a good starting value in the search for $\boldsymbol{\eta}^*$, if $\boldsymbol{\theta}^*$ is not too different from $\boldsymbol{\theta}$. In order to keep our analytical requirements to a minimum, we would like our method only to require the computation of the objective $\mathcal{L}(\boldsymbol{\eta}) = \prod_{t=1}^{N} f(\hat{\boldsymbol{y}}_t | \hat{\boldsymbol{x}}_{t-1}, \boldsymbol{\eta})$.

Given these considerations, to find the mle we run a Markov chain for $\boldsymbol{\eta}$ using the simulated data. Since our goal is to find the mle and the sample size is large, we use a

flat prior on $\boldsymbol{\eta}$ when running this chain. With the large sample size, the Markov chain will quickly move from the $\boldsymbol{\eta}$ to values near $\boldsymbol{\eta}^*$. We use a normal random walk Metropolis within Gibbs approach. That is, we first subdivide the $\boldsymbol{\eta}$ vector into subvectors. In the manner of a Gibbs sampler, we cycle through the subvectors one at a time. For subvector $\boldsymbol{\eta}_{(i)}$, we use the normal proposal $q(\boldsymbol{\eta}_{(i)}, \boldsymbol{\eta}_{(i)}^*) \sim n(\boldsymbol{\eta}_{(i)}, \boldsymbol{\Sigma}_{(i,i)})$ in a standard random walk Metropolis algorithm. Effectively, this is a simulated annealing optimization algorithm where the simulation size N is the temperature parameter because N is what controls the peakedness of the likelihood. A side benefit is that the chain for $\boldsymbol{\eta}$ also provides the scaling for the model assessment strategy proposed in Section 4.

We choose a fixed number of steps to run this chain, and keep the visited $\boldsymbol{\eta}$ which has the highest likelihood under the simulated data. In our experience, it is relatively straightforward to choose (i) a simulation sample size which is large enough to ensure that the map is adequately recovered by the mle and (ii) a number of steps to iterate the Markov chain in $\boldsymbol{\eta}$ that will ensure the chain has finished moving away from the starting value of $\boldsymbol{\eta}$.

This is a computationally costly part of our overall procedure. Since the simulation sample size N is large, each computation of the likelihood for the $\boldsymbol{\eta}$ chain can take a long time. Nonetheless, we have found that, because of the large N, this part of the procedure is remarkably stable, even though the statistical model may actually be difficult to estimate on data samples of the size n that we actually observe.

The main reason for placing $\boldsymbol{\theta}$ on a grid is that a significant reduction in computational time can be achieved. With $\boldsymbol{\theta}$ on a grid, it takes only a modest amount of memory to store all previously computed values of $\boldsymbol{\eta}$, $\mathcal{L}(\boldsymbol{\eta})$, $\pi(\boldsymbol{\theta})$, etc., in a binary tree indexed by $\boldsymbol{\theta}$. When $\boldsymbol{\theta}$ is revisited, both the fact that it is a revisit and the information required for Step 4 of the Metropolis algorithm for $\boldsymbol{\theta}$ can be quickly obtained by traversing the tree. The two most costly steps, 2 and 3, are thereby eliminated. By storing previous results in a tree and looking them up, the $\boldsymbol{\theta}$ chain runs faster as it becomes longer. We might also note that putting $\boldsymbol{\theta}$ on a grid reduces the accuracy to which $\boldsymbol{\eta}^*$ needs to be computed in order to compute α in Step 4 with sufficient accuracy.

In the supplemental materials for this article at the *JASA* web site are displayed the results of ten runs of an $\boldsymbol{\eta}$ chain (from Section 5). In the figure, one sees that the likelihood quickly increases as the $\boldsymbol{\eta}$ value moves toward the mle. The segments level off at different likelihoods because they represent the likelihoods of different simulated data sets. Because

of the large size of each simulated data set, $N = 50,000$ in this instance, the posterior is very tight around the mle and the chain quickly moves to a new level.

3.4   Identification and Map Recovery

The scientific model and the statistical model must work in concert with one another. In this subsection we discuss issues arising from the relationship between the two models.

The map $g(\theta) = \eta$ should not be one-to-many if Assumption 1 is to be satisfied. This is equivalent to stating that the statistical model should be identified by simulations from the scientific model. Identification of the statistical model by simulations from the scientific model entails some obvious conditions such as that the support of the statistical model should include the support of the scientific model. A violation of the support condition can occur if the scientific model has fewer random shocks than the dimension of y thereby causing the support of the scientific model to be a lower dimensional submanifold of the support of the statistical model. This can happen inadvertently if a proposed θ implies a singular variance matrix somewhere within the scientific model. However, singularity is easy to check.

Other violations can be more subtle. An example is the statistical model $y = \eta_1 \exp(\eta_2 x) + \eta_3 \exp(\eta_4 x) + e$ with simulated data from a scientific model that is actually $y = \theta_1 \exp(\theta_2 x) + e$ in disguise. In this case, either $\eta_3 = 0$ and η_4 is not identified or $\eta_2 = \eta_4$ and only the sum $\eta_1 + \eta_3$ is identified. One can usually detect this situation by checking the η chain described in Subsection 3.1 for a unit root. One way to do this is to find the eigenvector ℓ of the variance matrix of the chain $\{\eta_t\}_{t=1}^N$ with largest eigenvalue and examine the sequence of inner products $\{\ell' \eta_t\}_{t=1}^N$ for a unit root.

Lack of identification of the statistical model does not matter in the computation of the θ chain described in Subsection 3.1 as long as the likelihood is actually maximized at the computed η. The implied map $g(\theta) = \eta$ will be one-to-many but often considerations similar to estimability in less than full rank linear models come into play so that the features of the statistical model that are of interest in an application have the same value regardless of which maximizing value of η is chosen. These considerations are discussed in Gallant (1987). The methods proposed in Section 4 would have to be modified if η is not identified.

An omnibus check for identification failure and nearly anything else that can go wrong with the algorithm of Subsection 3.1, such as poor start values or not running the chain long enough to compute η^* to sufficient accuracy, is to run a regression of all computed

$\boldsymbol{\eta}^*$ on low degree polynomials in the corresponding values of $\boldsymbol{\theta}$. If the R^2 are high one has some assurance that the statistical model is identified and that accuracy is adequate. For the application described in Section 5, the 25%, 50%, and 75% quantiles of the R^2 for regressions of each component of $\boldsymbol{\eta}^*$ on a cubic in $\boldsymbol{\theta}$ are 0.91, 0.97, and 0.98.

Another requirement of Assumption 1 is that (1) should hold. To check directly whether (1) holds for a specific value of $\boldsymbol{\theta}$ one may (a) fit the statistical model to a simulation of the scientific model at that value of $\boldsymbol{\theta}$, (b) simulate from the fitted statistical model, and (c) check to see if the empirical distributions of the two simulations match. An illustration of this exercise is Figure 2 in the supplemental materials for this article at the *JASA* web site. This can serve as a partial empirical check if (1) cannot be established analytically.

The investigator may be motivated to use a statistical model that is known from experience to be in accord with the data. This can make the uncovered map more interpretable and easier to compute and facilitate the analysis in other ways as well. This could, however, lead to a violation of (1). The likelihood is then based on the statistical model closest in Kullback-Liebler divergence to the scientific model rather than the scientific model itself. Elaborations of the statistical model in order to satisfy (1) beyond those needed to fit the data may not have much effect on inference for $\boldsymbol{\theta}$ or the functionals of $p(\boldsymbol{y}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta})$ of interest. We see an example of this in Subsection 5.7. Intuition would suggest that this be true because it should make little difference if a feature is suppressed by the data in the fit or by the investigator before the fit.

Deleting a feature that is both in the data and in the scientific model is another matter. In Subsection 5.5, we see an instance where this sort of deletion changes results rather dramatically.

4   Inference Off the Manifold: Model Assessment

In this section $\boldsymbol{\eta} \in \boldsymbol{H}$ becomes the parameter of interest. The scientific model $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ may be viewed as a sharp prior that restricts $\boldsymbol{\eta}$ to lie on the manifold $\boldsymbol{\mathcal{M}} \subset \boldsymbol{H}$. What one would like to do is to see how results change as this prior is relaxed. However, once we have moved off the manifold $\boldsymbol{\mathcal{M}}$ we can no longer view results from the perspective of the scientific model $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ and must view them from the perspective of the statistical model $f(\boldsymbol{y}|\boldsymbol{x}|, \boldsymbol{\eta})$ because the scientific model loses meaning off the manifold. Therefore, seeing how

results change must be taken to mean seeing how the marginal posterior distribution of a parameter or functional of the statistical model changes. Denote the vector of functionals of the statistical model of interest by

$$\Upsilon : f(\cdot|\cdot, \boldsymbol{\eta}) \mapsto \boldsymbol{v}. \tag{5}$$

For convenience, if an element of $\boldsymbol{\eta}$ is of interest, we make it an element of $\boldsymbol{v}$.

We assume that we have a discrete set of points on the manifold $\{\boldsymbol{\eta}_j \in \mathcal{M} : j = 1, \ldots, G\}$. The analysis of Section 3 generates a discrete set of points $\{\boldsymbol{\theta}_j \in \Theta : j, \ldots, G\}$ at which the map $\boldsymbol{g}$ has been evaluated; putting $\boldsymbol{\eta}_j = \boldsymbol{g}(\boldsymbol{\theta}_j)$ provides such a set of points. Relaxation of the prior will be formulated in terms of a weighted distance of $\boldsymbol{\eta}$ from the manifold $\mathcal{M}$. We can cheaply compute the distance from $\boldsymbol{\eta}$ to the manifold as

$$d(\boldsymbol{\eta}, \mathcal{M}) = \min_{j=1,\ldots,G} (\boldsymbol{\eta} - \boldsymbol{\eta}_j)' \boldsymbol{A}_j (\boldsymbol{\eta} - \boldsymbol{\eta}_j), \tag{6}$$

where the $\boldsymbol{A}_j$ are scaling matrices. Let $\hat{j}$ denote the index j at which the minimum occurs, let $\hat{J}$ denote the map

$$\hat{J} : \boldsymbol{\eta} \mapsto \hat{j}, \tag{7}$$

and let $\hat{\boldsymbol{h}}(\boldsymbol{\eta}) = \boldsymbol{\eta}_{\hat{J}(\boldsymbol{\eta})}$.

The prior we propose is the product of preferences along the manifold, preferences about how close the statistical model is to the manifold, and general preferences about $\boldsymbol{\eta}$,

$$\pi_\kappa(\boldsymbol{\eta}) \propto w_1[\hat{\boldsymbol{h}}(\boldsymbol{\eta})] \exp\left(-\frac{d(\boldsymbol{\eta}, \mathcal{M})}{2\kappa}\right) w_3(\boldsymbol{\eta}), \tag{8}$$

where $w_1(\boldsymbol{\eta})$ and $w_3(\boldsymbol{\eta})$ are suitably chosen positive functions and we assume the middle $w_2(\boldsymbol{\eta}, \kappa)$ term assures integrability. The three terms in the product (8) correspond to our three kinds of preferences regarding $\boldsymbol{\eta}$. The prior becomes more diffuse and the scientific model less influential as the scale factor κ increases. Note that none of the individual terms is thought of as being a prior in its own right, each is just a part of the overall construction. To check that (8) results in a reasonable prior, draws of $\boldsymbol{\eta}$ may be simulated from the prior and prior marginals of interest checked (see Figure 5).

The functions $w_1(\boldsymbol{\eta})$ and $w_3(\boldsymbol{\eta})$ in (8) may be related to the preferences along $\mathcal{M}$ that we had for $\boldsymbol{\theta}$ in the scientific model, but there is no logical necessity that this be the case. If we desire to use roughly the same kind of preferences along $\mathcal{M}$ for $\boldsymbol{\eta}$ as we used for $\boldsymbol{\theta}$, we

can use $\boldsymbol{\eta}_j = \boldsymbol{g}(\boldsymbol{\theta}_j)$ to generate our points on the manifold, put

$$w_1[\hat{\boldsymbol{h}}(\boldsymbol{\eta})] \propto \pi(\boldsymbol{\theta}_{\hat{J}(\boldsymbol{\eta})}), \tag{9}$$

where $\pi(\boldsymbol{\theta})$ is given by (3) and $\hat{J}(\boldsymbol{\eta})$ by (7), and put $w_3(\boldsymbol{\eta}) = 1$. Note that $\pi(\boldsymbol{\theta}_{\hat{J}(\boldsymbol{\eta})})$ is a composite function that depends only on $\boldsymbol{\eta}$ that is easily retrieved from our stored map. With these choices, if $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ both have the same distance from the manifold, then $\pi_\kappa(\boldsymbol{\eta}_1)/\pi_\kappa(\boldsymbol{\eta}_2) = \pi(\boldsymbol{\theta}_{\hat{J}(\boldsymbol{\eta}_1)})/\pi(\boldsymbol{\theta}_{\hat{J}(\boldsymbol{\eta}_2)})$. Recall that one of our motivating considerations was the problem of sparse data and that this problem is overcome by the introduction of prior information. A small κ may imply sufficient prior information for inference. With large κ, additional prior information may be needed in the form of a choice of $w_3(\boldsymbol{\eta})$.

The computation of the posterior distribution of $\boldsymbol{\eta}$ using the statistical model $f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\eta})$ and prior $\pi_\kappa(\boldsymbol{\eta})$ can be accomplished by a routine application of the Metropolis algorithm because $\pi_\kappa(\boldsymbol{\eta})$ is easily computable and an analytic expression for $f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\eta})$ is available.

Our proposal is that the scientific model be assessed by plotting a suitable measure of the location and scale of the posterior distribution of $\boldsymbol{v}$ against κ or, better, sequential density plots as in Figure 5. What one expects to see, for a well fitting scientific model, is that the location measure does not move by a scientifically meaningful amount as κ increases, which indicates that the model fits, and that the scale measure increases, which indicates that the scientific model has empirical content. What we see in Figure 5 is that as κ increases the scale of two functionals of interest increases, indicating empirical content, but that location also shifts, providing evidence against the model.

For the scaling matrices $\boldsymbol{A}_j$, we put $\boldsymbol{A}_1 = \ldots = \boldsymbol{A}_G = \boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{-1}$ in (6) where $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ is computed as follows: Initialize to zero. Whenever the Metropolis-Hastings chain for computing the mle of $\boldsymbol{\eta}$ at Step 3 of the Metropolis algorithm in Subsection 3.1 must be run, update

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}} \leftarrow \boldsymbol{\Sigma}_{\boldsymbol{\eta}} + (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)' \tag{10}$$

where $\boldsymbol{\eta}_1$ is a point on the chain immediately after transients have died out and $\boldsymbol{\eta}_2$ is the last point on the chain. This method of scaling the distance measure is reasonable because it puts $\boldsymbol{\eta}$ on the scale of the posterior: Distance is being measured in units of standard deviation.

The distance function (6) can be made invariant to a linear reparameterization of the statistical model and first order invariant to a nonlinear transformation by letting $\boldsymbol{A}_j$ be

the inverse of the covariance of the draws $\boldsymbol{\eta}_t^j$ from the chain used to obtain $\boldsymbol{\eta}_j = \boldsymbol{g}(\boldsymbol{\theta}_j)$ at Step 3 of the Metropolis algorithm. The scaling proposed in the paragraph above will achieve approximate invariance if the $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_j}$ are relatively homogeneous.

One should note that using posterior draws from the scientific model to compute the image $\boldsymbol{\mathcal{M}}$ of the map $\boldsymbol{g}(\cdot)$ does make use of the data to determine the prior $\pi_\kappa(\boldsymbol{\eta})$. We do not regard this as a problem because the $\boldsymbol{\theta}$ draws will contain enough extreme values to make sure that the extent of $\boldsymbol{\mathcal{M}}$ is large enough. If one is particularly worried about this, one could run the $\boldsymbol{\theta}$ chain with a smaller amount of data to make sure that $\boldsymbol{\mathcal{M}}$ is over explored.

5   Habit Persistence Asset Pricing Model

In this section we shall apply the proposed methods to the habit persistence asset pricing model of Campbell and Cochrane (1999), referred to as CC hereafter. Although it is widely viewed as a behavioral model and it is the result of an admitted attempt to reverse engineer away the statistical inadequacies of the classical consumption based asset pricing model (Lucus 1978), the habit model actually can be justified from plausible micro-foundations (Guvenen 2005). The habit model exhibits all the characteristics discussed in Section 1: (1) the likelihood is not available; (2) prior information on model parameters is available; (3) prior information in the form of restrictions on model functionals is available; (4) the model can be simulated; (5) a generally accepted statistical model for its data is available.

In the remainder of this section, we describe the data, introduce the habit model, and apply the methods that we have proposed. We conclude that the habit model is not supported by the data using indirect evidence, Figure 5, and a direct statistical test.

We also find, Table 1, that statistical methods that rely on a counterfactual assumption that the data are conditionally homoskedastic generate reasonable parameter estimates. However, when the constraint of homoskedasticiy is removed, the parameter estimates become implausible. From the implied map we discover in Figure 4 that a sharply delineated subset of the habit model's parameters control the observable conditional heteroskedasticity. This subset contains the parameters that describe preferences of which the risk aversion parameter is particularly important. BGT argue that the habit model is not supported by the data using indirect evidence that differs from ours but their direct frequentist test accepts the habit model. Our results suggest that BGT's failure to formally reject is a consequence

of using statistical methods that impose conditional homoskedasticity under both the null and the alternative. However, without the use of prior information, one is nearly forced to impose conditional homoskedasticity and other undesirable restrictions on the alternative due to the sparseness of the data. Commentary in BGT speaks to this issue.

5.1 The Data

The data are per capita non-durables and services consumption C_t^a over year t and the price P_{dt}^a at end of year t of a value weighted portfolio comprised of all stocks listed on the New York and American exchanges. The observation period is 1929–2001. Both series are adjusted for inflation. Data sources and collection protocol are described in BGT. The corresponding annual consumption growth series and stock returns series are $\Delta c_t^a = \log(C_t^a) - \log(C_{t-1}^a)$ and $r_{dt}^a = \log(P_{dt}^a) - \log(P_{d,t-1}^a)$. In our previous vector notation, we have $\boldsymbol{y}_t = (\Delta c_t^a, r_{dt}^a)'$, $t = 1, \ldots, n = 72$. Annual data is used rather than monthly or quarterly data because seasonality issues are avoided and the historical record is longer.

We next discuss how to simulate the habit model. The model is simulated at the monthly frequency and then aggregated to the annual frequency. This is consistent with CC and BGT and is common in this literature.

5.2 Model Description

The intuitive notions behind any consumption based asset pricing model are that agents receive wage income and dividend income from which they purchase consumption. Agents seek to reallocate their consumption over time by trading shares of stock that pay a random dividend and bonds that pay interest with certainty. The driving processes of such a model are wages and dividends or, equivalently, consumption and dividends because wages can be recovered by subtracting dividends from consumption due to the fact that someone must own the stock so the dividends must be received while for bonds someone pays interest and another receives so there are no net bond receipts. Agents are endowed with a utility function that depends on the entire consumption process. The parameters of this function determine their preferences. The first order conditions of their utility maximization problem determine the prices at which they are willing to trade securities. We shall describe the driving processes, the utility function, and the first order conditions of the habit model, in that order.

The driving processes of the habit model are consumption and dividend growth

$$c_t - c_{t-1} = g + v_t, \tag{11}$$

$$d_t - d_{t-1} = g + w_t.$$

We follow the standard convention that lower case denotes the logarithm of an upper case variable; e.g. $c_t = \log(C_t)$, $d_t = \log(D_t)$. The errors (v_t, w_t) in (11) are normal with mean zero and variance $\mathrm{Var}\,(v_t, w_t) = \boldsymbol{R}\boldsymbol{R}'$, where $\boldsymbol{R}$ is upper triangular with nonzero elements r_{11}, r_{12}, and r_{22}. At times it is more convenient to express the variance matrix in terms of $\sigma^2 = \mathrm{Var}(v_t)$, $\sigma_w^2 = \mathrm{Var}(w_t)$, and $\rho = \mathrm{corr}(v_t, w_t)$.

Upon exponentiation to get C_t and summing over consecutive blocks of twelve, the consumption process does correspond conceptually to the data series C_t^a described above. This operation puts first order serial correlation in the annual simulation; first order serial correlations is in the data also. On the other hand, the dividend process does not correspond conceptually to observable data primarily because what can be observed is strongly influenced by tax policy causing, e.g., corporations to shift dividend payments into or out of stock repurchases. Therefore, d_t is to be regarded as a latent variable. It remains to consider how the returns process r_{dt}^a is simulated.

The habit model asserts that all agents in the economy are endowed with the same utility function $\mathcal{E}_0 \sum_{t=0}^{\infty} \delta^t [(C_t - X_t)^{1-\gamma} - 1]/(1 - \gamma)$ where X_t is habit, δ the time discount factor, γ is the risk aversion parameter, and $\mathcal{E}_t$ is conditional expectation given past values of C_t, D_t, and X_t. Habit is determined by $X_t = C_t - S_t C_t$, where S_t, called the surplus ratio, is the state variable of the model and is designed to behave as if it followed a discretely sampled conditionally heteroskedastic diffusion. Specifically, and recalling the upper and lower case convention,

$$s_t - \bar{s} = \phi(s_{t-1} - \bar{s}) + \lambda(s_{t-1})v_t, \tag{12}$$

where $\lambda(s) = \frac{1}{S}\sqrt{[1 - 2(s - \bar{s})]} - 1$ if $s_t \leq s_{\max}$ and zero else; $\bar{s}$ and $s_{\max}$ are computed from model parameters $\boldsymbol{\theta} = (g, r_{11}, r_{12}, r_{22}, \phi, \delta, \gamma)$ as $\bar{S} = \sqrt{[(r_{11}^2 + r_{12}^2)(\gamma/(1 - \phi))]}$ and $s_{\max} = \bar{s} + (1 - \bar{S}^2)/2$.

Agents are presumed to be so numerous that each can solve their own utility maximization problem without regard to the actions of the others. Under this assumption, the price dividend ratio satisfies $P_{dt}/D_t = V_{\boldsymbol{\theta}}(S_t)$, where $V_{\boldsymbol{\theta}}(\cdot)$ solves the integral equation

$$V_{\boldsymbol{\theta}}(S_t) = \mathcal{E}_t \left\{ \delta \left(\frac{S_{t+1}C_{t+1}}{S_t C_t} \right)^{-\gamma} \left(\frac{D_{t+1}}{D_t} \right) [1 + V_{\boldsymbol{\theta}}(S_{t+1})] \right\}. \tag{13}$$

16

These are the first order conditions of each agent's optimization problem. The solution method we use is described in BGT. Two aspects of the solution method are relevant to us: the method uses a long simulation of (C_t, D_t, S_t) to approximate an integral by an average and uses a nonlinear equation solver that needs starting values to compute $V_{\boldsymbol{\theta}}(\cdot)$. An important consequence is that if two adjacent draws $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_{i+1}$ in an MCMC chain are close together then the answer left over from solving for draw i makes a good starting value for solving for draw $i + 1$.

What one does, then, for $\boldsymbol{\theta}$ given, is first simulate consumption and dividends using (11) and compute the surplus ratio from them using (12) to get the simulation $\{\hat{C}_t, \hat{D}_t, \hat{S}_t\}_{t=1}^{12N}$, where $12N$ is to have a simulation of length N after aggregation. Second, use the simulation and the BGT method to solve (13) for $V_{\boldsymbol{\theta}}(\cdot)$. Third, compute stock price as $\hat{P}_{dt} = \hat{D}_t V_{\boldsymbol{\theta}}(\hat{S}_t)$. Fourth, compute geometric stock returns using $\hat{r}_{dt} = \log(\hat{P}_{dt}) - \log(\hat{P}_{d,t-1})$. A computation similar to the second through third steps provides the geometric risk free interest rate $\hat{r}_{ft}$. Fifth, compute functionals of interest from the simulation. An important functional is $\mathcal{E}(r_{ft}^a)$ computed by averaging r_{ft} over consecutive blocks of twelve. Last, aggregate $\{\hat{C}_t, \hat{r}_{dt}\}_{t=1}^{12N}$ by summing consecutive blocks of twelve to get the annual series $\{\hat{C}_t^a, \hat{r}_{dt}^a\}_{t=1}^{N}$ and put $\hat{\boldsymbol{y}}_t = (\log \hat{C}_t^a - \log \hat{C}_{t-1}^a, \hat{r}_{dt}^a)$.

5.3   Prior Information

Our discretization of $\boldsymbol{\theta}$ (see Subsection 3.2) automatically restricts our prior support to the set of $\boldsymbol{\theta}$ such that $a_i \le \theta_i \le b_i$ for each i. From the marginal posteriors (Figure 3) we can see that our choices of a_i and b_i are not influential since none of the distributions appear to be truncated.

A more complex support condition derives from the fact that a solution $V_{\boldsymbol{\theta}}(\cdot)$ to (13) does not exist for all $\boldsymbol{\theta}$. We view this as an indication that such a $\boldsymbol{\theta}$ is unreasonable and assign it zero prior probability. To protect against spurious failure to solve (13) we rely on small proposals for $\boldsymbol{\theta}$ and a long list of recent solutions to (13) from which we choose the most promising start. If we fail to solve (13) we set the prior $\pi(\boldsymbol{\theta})$ equal to zero. In our Metropolis algorithm (Subsection 3.1), this determination is made at step 2. In the case of failure we set α to zero and proceed directly to step 5.

The rest of our prior information will be expressed through $\psi(\boldsymbol{\theta}) = \mathcal{E}(r_{ft}^a)$ (a functional of $p(\cdot|\cdot, \boldsymbol{\theta})$, computed from the simulation as described above), the correlation between monthly

consumption and dividend growth $\rho(\boldsymbol{\theta})$ (a simple function of $\boldsymbol{\theta}$), and ϕ (an element of $\boldsymbol{\theta}$). In the notation of (3), Section 2, we must define a function $h(\boldsymbol{\theta}, \psi(\boldsymbol{\theta})) \propto \pi(\boldsymbol{\theta})$ that, at least approximately, captures these preferences.

In thinking about the risk free rate we first note that it is not directly observable because its computation from observed interest rates requires estimation of anticipated inflation (Mishkin 1981). Because the evidence indicates that the risk free rate is low with a variance that is much smaller than the variance of inflation (Campbell 2003), an attempt to generate a risk free rate series over our sample period would produce a series that would consist almost entirely of measurement error. We base our prior information on the values that Campbell determined from several long historical time series. We put most of our prior mass on $\boldsymbol{\theta}$ such that $\psi(\boldsymbol{\theta}) = \mathcal{E}(r^a_{ft})$ is in the vicinity of .89%.

A consequence of treating dividends as latent is that $\rho(\boldsymbol{\theta})$ is poorly determined by the observables. Thus, prior information is important. The parameter ϕ is an autoregressive parameter and we find that it is necessary to prevent the MCMC chain from putting ϕ too close to one and yet our prior belief is that it is very close to one. We put prior mass on $\boldsymbol{\theta}$ such that $\rho(\boldsymbol{\theta})$ is close to .2 and ϕ is very close to .9884. The values are based on CC. Note that in terms of $\boldsymbol{\theta}$, $\rho(\boldsymbol{\theta}) = r_{12}/\sqrt{(r_{12}^2 + r_{22}^2)}$.

We let

$$\pi(\boldsymbol{\theta}) \propto e^{-\frac{1}{2}\left(\frac{\psi(\boldsymbol{\theta})-.89}{0.5}\right)^2} e^{-\frac{1}{2}\left(\frac{\rho(\boldsymbol{\theta})-.2}{0.05}\right)^2} e^{-\frac{1}{2}\left(\frac{\phi-0.9884}{0.005}\right)^2} I(\boldsymbol{\theta}) \tag{14}$$

where $I(\boldsymbol{\theta})$ is an indicator function capturing the support conditions discussed above.

Although we use the normal kernel in our construction, nothing is normally distributed. For example, the reader is reminded that we discretize $\boldsymbol{\theta}$ so that its support lies on a finite number of atoms. We use the normal kernel as a simple way of making sure that the resulting prior $\pi(\boldsymbol{\theta})$ is relatively small for $\boldsymbol{\theta}$ corresponding to ψ, ρ, and ϕ values which are "too far" from the values .89, .2, and .9884. The scaling quantities .5, .05, and .005 used above help determine what "too far" is. Certainly, none of the normal kernels in the expression above represents a prior in its own right. Each is merely a part of the overall construction just as in the case of the indicator function.

The inclusion of the middle term involving $\rho(\boldsymbol{\theta})$ creates prior dependence between r_{12} and r_{22}. Although we have no analytical knowledge of the function $\psi(\boldsymbol{\theta})$, CC's study of numerical solutions suggest that it is a complex non-linear function involving all components of $\boldsymbol{\theta}$. Thus, inclusion of the first term captures rich and complex prior information, involving

every element of $\boldsymbol{\theta}$, that would be very hard to impose in any other way.

<center>Figure 2 about here</center>

In order to see if our construction results in an acceptable prior, we simulate from the prior and report the resulting marginal draws of each θ_i as well as ψ and ρ. Drawing from the prior distribution and then looking at important functions of $\boldsymbol{\theta}$ to deduce what the prior really says is useful in many Bayesian analyses (see Geweke (2005, section 8.3)). Figure 2 shows the marginal priors. The histograms in the first two rows and (3,1) position correspond to the seven elements of $\boldsymbol{\theta}$ while the histograms in the (3,2) and (3,3) positions depict the marginal priors for $\psi(\boldsymbol{\theta}) = \mathcal{E}(r_{ft}^a)$ and ρ.

All marginals are acceptably in accordance with our prior preferences. The ϕ is probably quite close to one but bounded away from one (the grid limits are $(a_5, b_5) = (.9, .99)$). The $\mathcal{E}(r_{ft}^a)$ is, as a practical matter, in the vicinity of .89%. The prior for ρ gives support to values close to .2. The marginal for ρ is somewhat surprising in that extension of the right tail out towards one is not at all evident from our construction. This must be due to the information provided through ψ. We could tighten up the prior for ρ by adjusting the scale factor in the middle term of our expression of π, but we find the current result acceptably informative. Similarly, the prior information on δ comes through ψ. If we compare the marginal priors (Figure 2) to the marginal posteriors (Figure 3), we see that the data provides substantial information (except, perhaps, for ϕ, where our prior is very strong).

Again, we emphasize that while we use normal densities in our construction, this does not imply that the corresponding marginals are normal. While the prior marginal for ψ does seem to be in accordance with the normal in the first term in our construction, the marginal prior for ρ certainly is not!

5.4   The Statistical Model

The statistical model is a bivariate GARCH system:

$$
\begin{pmatrix} c_t^a - c_{t-1}^a \\ r_{dt}^a \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \tag{15}
$$
$$
+ \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} c_{t-1}^a - c_{t-2}^a - b_1 \\ r_{d,t-1}^a - b_2 \end{pmatrix}
$$

<center>19</center>

$$+ \begin{pmatrix} r_{11,t-1} & r_{12,t-1} \\ 0 & r_{22,t-1} \end{pmatrix} \begin{pmatrix} z_{1t} \\ z_{2t} \end{pmatrix}$$

$$\begin{pmatrix} r_{11,t-1} \\ r_{12,t-1} \\ r_{22,t-1} \end{pmatrix} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \end{pmatrix} \tag{16}$$

$$+ \begin{pmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \\ \rho_{31} & \rho_{32} \end{pmatrix} \begin{pmatrix} |z_{1,t-2}| \\ |z_{2,t-2}| \end{pmatrix}$$

$$+ \tau \begin{pmatrix} r_{11,t-2} - \rho_1 \\ r_{12,t-2} - \rho_2 \\ r_{22,t-2} - \rho_3 \end{pmatrix}$$

where (z_{1t}, z_{2t}) are iid. The statistical model $f(\boldsymbol{y}_t | \boldsymbol{x}_{t-1}, \boldsymbol{\eta})$ has sixteen parameters $\boldsymbol{\eta} = (b_1, b_2, b_{11}, b_{21}, b_{12}, b_{22}, \rho_1, \rho_2, \rho_3, \rho_{11}, \rho_{21}, \rho_{31}, \rho_{12}, \rho_{22}, \rho_{32}, \tau)$.

This is a standard type of model for the observed annual data (Tauchen and Hussey 1991). CC and BGT suppress the GARCH terms in (16); i.e., all but ρ_1, ρ_2, and ρ_3 in (16) are put to zero. In order to assess the effect of suppression, we consider both the case when GARCH is present and absent. When the GARCH terms are absent, we call the statistical model a VAR.

One must remember that these are bivariate, annual data on consumption growth and stock returns. While the BIC criterion indicates that a fat-tailed conditional density is required for univariate, daily stock returns (Chernov, Gallant, Ghysels, and Tauchen 2003), it does not indicate that fat tails are required for bivariate, annual consumption growth and stock returns. The BGT analysis of the habit model assumes normality as implicitly does the Campbell and Cochrane (1999) calibration as indicated by their choice of which moments to match. These considerations and a desire to be able to compare with the results of others lead us to adopt normal innovations as our preferred statistical model.

The above considerations relate to data. Notwithstanding, even after aggregation to the annual frequency, the habit model behaves more like a model for high frequency data than a model for annual data. If one applies the methods for finding an auxiliary model

20

proposed by Chernov, Gallant, Ghysels, and Tauchen (2003) to a simulation from the habit model at the parameters shown in column Bayes-GARCH in Table 1, one arrives at a model with a two-lag nonlinear mean function and fat-tailed GARCH innovations. In our view, a two-lag nonlinear mean function is too counterfactual for annual data to be taken seriously. Therefore we do not further modify the statistical model to put additional structure in the mean function. Fat-tailed innovations may be a different matter.

In the supplemental materials for this article at the *JASA* web site is the result of an application of the diagnostic technique proposed in Subsection 3.4 where the statistical model is fitted to simulations from the scientific model and then itself simulated. Densities estimated from these two simulations are then over-plotted. These plots indicate that the habit model does put fat tails into the annual returns series and that the tails of the returns innovation in the statistical model must be like a *t*-distribution on 3.5 degrees of freedom to match the simulation of the habit model.

To summarize, our preferred statistical model is the GARCH model given by (15) and (16) with normal errors. We do not consider an extension of our preferred model to the case where the one-lag linear mean function in (15) is replaced by a two-lag nonlinear mean function. We do consider the case where normal errors are replaced by fat-tailed errors in Subsection 5.7 and the case where the GARCH structure (16) is suppressed in Subsection 5.5. In Poirier's (1998) terminology, these different statistical models define different windows through which the world is viewed and trying several views is a sensitivity analysis that determines how much difference a different view might make.

5.5 Model Estimation: Scientific Model Imposed

We ran the $\boldsymbol{\theta}$ chain for 800,000 iterations, keeping every 8th leaving 100,000. The chain was started near the mode of the posterior density to avoid problems with transients. The mode was determined from several initial runs of size 100,000. Visual inspection of time series plots of the draws (not shown) indicated that this strategy for eliminating transients was successful.

<center>Table 1 about here</center>

Table 1 reports our results for the VAR and GARCH statistical models together with BGT estimates. Estimates for the model as described are shown in the top half of the table. These estimates are for the monthly frequency. For convenience in interpretation,

<center>21</center>

annualized estimates are displayed in the bottom half. Further down are estimates for $\mathcal{E}(r_{f,t}^a), \sqrt{\mathrm{Var}(r_{f,t}^a)}, \mathcal{E}(r_{d,t}^a), \sqrt{\mathrm{Var}(r_{d,t}^a)}$.

In the columns of Table 1 that are labeled mode, the values shown are the mode of the multivariate posterior for $\boldsymbol{\theta} = (g, r_{11}, r_{21}, r_{22}, \phi, \delta, \gamma)$, not the mode of each marginal. The values of the multivariate mode appear in the MCMC chain for $\boldsymbol{\theta}$ and therefore do correspond to an actual simulation of the habit model whereas the vector of marginal modes or means might not (and may not even satisfy support conditions). The values for $\mathcal{E}(r_{f,t}^a)$, $\sqrt{\mathrm{Var}(r_{f,t}^a)}, \mathcal{E}(r_{d,t}^a), \sqrt{\mathrm{Var}(r_{d,t}^a)}$ in the columns labeled mode are computed from the mode of $\boldsymbol{\theta}$. Those in the columns labeled mean are the means of the marginal posterior distribution of $\mathcal{E}(r_{f,t}^a), \sqrt{\mathrm{Var}(r_{f,t}^a)}, \mathcal{E}(r_{d,t}^a), \sqrt{\mathrm{Var}(r_{d,t}^a)}$. All standard deviations in the columns headed Bayes are computed from marginal posteriors. The BGT estimates optimize a criterion function and therefore correspond to a simulation; the values shown for $\mathcal{E}(r_{f,t}^a), \sqrt{\mathrm{Var}(r_{f,t}^a)}, \mathcal{E}(r_{d,t}^a)$, $\sqrt{\mathrm{Var}(r_{d,t}^a)}$ are computed from that simulation.

As with any discrete time dynamic model, annualized parameter values for the habit model do not have the property that simulations obtained by running the model at the annual frequency using the annualized parameter values will have the same distribution as simulations obtained by aggregating a monthly simulation. For this reason, the values at the monthly frequency in the upper half of of Table 1 should be regarded as the definitive estimates of model parameters. The annualized values in the lower half of the table are to be regarded only as an aid in their interpretation.

<div align="center">Figure 3 about here</div>

The Bayes-GARCH columns of Table 1 report computations that use the GARCH statistical model. Their most striking feature is the large deviation of the Bayes-GARCH estimate of mean stock returns $\mathcal{E}(r_{d,t}^a) \doteq 11\%$ from the value computed from the data of 6.02% and from the BGT estimate of 6.54%. The reason for this is that the BGT estimator completely ignores the conditional heteroskedasticity in the data. The auxiliary model of the BGT estimator, whose impact on the BGT estimate is analogous to the impact of the statistical model on the Bayes procedure we propose, is a conditionally homoskedastic VAR. If we too use a conditionally homoskedastic VAR to implement our Bayes estimator, the results in the column labeled Bayes-VAR are obtained, bringing the Bayes estimate of $\mathcal{E}(r_{d,t}^a)$ into agreement with the data and with the BGT estimates. The Bayes-VAR estimates are also in good agreement with the values reported in CC that were determined by matching to a set

<div align="center">22</div>

of moments that are not rich enough to identify models with conditional heteroskedasticity.

The reasons that BGT had to suppress GARCH in their auxiliary model are twofold. Observable data rather than prior information are used to achieve identification in BGT thereby creating the need for a four dimensional series. The data ($n = 72$) are too sparse to support a GARCH specification in four dimensions. We, on the other hand, have no such difficulties because we are fitting to large simulations, not to data.

It is of interest to view the effect that using the VAR statistical model has on marginal posteriors. These are shown in Figure 3. Elimination of the requirement that the habit model confront the conditional heteroskedasticity in the data results in large shifts in the marginal posterior of the utility parameters (ϕ, γ, δ). This accounts for the dramatic left shift in the marginal posterior for $\mathcal{E}(r_{d,t})$ and dramatic variance reduction in the posterior for $\sqrt{\mathrm{Var}(r_d)}$ that is seen in Table 1. This, following the logic of Section 4, can be regarded as strong evidence against the VAR specification of the statistical model.

<center>Figure 4 about here</center>

It is also of interest to note that it is the preference parameters (ϕ, δ, γ) of the scientific model that control the GARCH parameters $(\rho_{11}, \rho_{21}, \rho_{31}, \rho_{12}, \rho_{22}, \rho_{32}, \tau)$ of the statistical model with γ having by far the most influence. This can be discovered by inspecting the map $g : \theta \mapsto \eta$. Since this map is high-dimensional we focus on the conditional correlation between annual stock returns and consumption growth for the year 2002. This correlation is important substantively because a high positive correlation implies that risk averse investors will require high expected returns to induce them to invest. The conditional correlation is a functional of the statistical model $f(\cdot|\cdot, \eta)$ determined by the map $\eta = g(\theta)$ and computed by applying the GARCH recursion (16) to the data. The last variance matrix given by the recursion is the conditional variance for the year 2002. In Figure 4 we plot the conditional correlation against the parameters $\theta = (g, r_{11}, r_{21}, r_{22}, \phi, \delta, \gamma)$ of the scientific model. The dots show the conditional correlation when the statistical model is GARCH and the circles when VAR. The difference between these two curves is the GARCH effect. In each panel, a θ_i is varied and the remaining elements of θ held fixed at their posterior means. The solid vertical line is the posterior mean of θ_i when the statistical model is GARCH and the dashed line when VAR. The point where the solid line crosses the dots gives the conditional correlation when $\eta = g(\theta)$ is evaluated at the posterior mean under GARCH; similarly the point where the dashed line crosses the circles gives the conditional correlation under VAR.

<center>23</center>

5.6 Model Assessment: Scientific Model Relaxed

We now apply the methodology described in Section 4 to the habit model. We construct a prior along the lines suggested is Section 4 as follows.

The first term of (8) uses the construction (9) of the form $w_1[\hat{\boldsymbol{h}}(\boldsymbol{\eta})] \propto \pi(\boldsymbol{\theta}_{\hat{J}(\boldsymbol{\eta})})$ with $\pi(\boldsymbol{\theta})$ given by (14). For the second term we use the scaling matrix (10) and the values 1, 20, and 100 for κ. As to the third term, for small κ the influence of the prior $\pi_\kappa(\boldsymbol{\eta})$ is strong enough that the MCMC chain for η behaves well with $w_3(\boldsymbol{\eta}) = 1$. But when κ is large, the parameter τ that appears in (16) can move to explosive values and stick. To fix this problem we set $w_3(\boldsymbol{\eta}) \propto \exp[-\frac{1}{2}\left(\frac{\tau-0.6}{0.08}\right)^2]$ for all κ.

The functionals of interest are the mapping Υ_1 of $f(\cdot|\cdot, \boldsymbol{\eta})$ to the mean return on the stock portfolio over the period 2002–2102 and the mapping Υ_2 to the conditional correlation between the return on the stock portfolio and consumption growth for the year 2002. The conditioning event for both functionals is the 73 years of observed data. The Υ_1 is computed from a realization obtained by simulating the GARCH model over the period 2002–2102; the Υ_2 is obtained from the variance matrix for the year 2002 computed as described at the end of Subsection 5.5. Both depend on the data and $\boldsymbol{\eta}$; the Υ_1 also depends on an initial seed that was the same for each $\boldsymbol{\eta}$.

<center>Figure 5 about here</center>

Using the methods proposed in Section 4, we computed MCMC chains for three values of κ (1,20,100). The stationary distribution of each chain is the posterior for $\boldsymbol{\eta}$ under prior κ. We then evaluated the two functionals (Υ_1, Υ_2) at each $\boldsymbol{\eta}$ in the chain and plotted their densities in Figure 5. The κ increases as we go down the rows. The left hand panels show prior (dashed lines) and posterior densities (solid lines) for the mean return and the right hand panels display the same for the correlation. For small κ, we have a tight prior and the prior and posterior are similar.

For larger κ, the posterior shifts. In the three right hand plots, the posterior distribution of the correlation shifts as κ increases while the prior remains reasonable. As κ increases we place less weight on $\boldsymbol{\eta}$ close to the scientific model. This is simple, interpretable evidence that the scientific model has trouble tracking a feature of the data. This feature is important substantively because a high positive correlation implies that risk averse investors will require high expected returns to induce them to invest.

<center>24</center>

The mean stock return, unlike a correlation, has a substantive meaning without having an independent statistical meaning. Nonetheless, coping with retirement plan options and the laws governing bequests has the made the notion of a mean return over a long planning horizon meaningful to most of us. For the mean return, we also see that as κ increases, the distribution shifts. The posterior mean of the mean return over the hundred year horizon is 0.1 in the top left panel and 0.08 in the bottom left; the standard deviation increases from 0.0127 to 0.0172. The location shift may not appear as dramatic as for the correlation functional but substantively it is quite large. This may be seen by noting that $1,000 of the stock would be worth $22,026,000 at the end of the period at the former rate vs. $2,980,000 at the latter.

Figure 5 focuses on the marginal priors and posteriors of two particular functionals of interest. Of course, as κ changes, the entire sixteen dimensional prior and posterior of $\boldsymbol{\eta}$ is changing. To get a sense of how the posterior is moving, we consider κ equal 1, 20, or 100 as a choice of model and compute the posterior probability for the three different models with each having prior probability 1/3. That is, the pair $(f(\cdot|\cdot, \boldsymbol{\eta}), \pi_\kappa(\boldsymbol{\eta}))$ is considered to be a model and the posterior probability of each κ choice is proportional to $\int f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\eta})\, \pi_\kappa(\boldsymbol{\eta})\, d\boldsymbol{\eta}$. The integrals were computed using method f_5 from Gamerman and Lopes (2006, section 7.2.1). Method f_5 requires draws from both the prior and the posterior but does not require the normalizing constant for the prior. We find that the posterior probability that κ equals 100 is 0.999. As we loosen up the prior and allow the posterior to move away from the scientific manifold, it is able to find much higher likelihoods. If we include the scientific model ($\kappa = 0$) in our list of models and use prior probabilities 1/4, the posterior probability that κ equals 100 is 0.996.

5.7   Sensitivity Analysis

To check the sensitivity of our results to the normality assumption in (15), we recomputed the columns labeled Bayes-GARCH in Table 1 using a statistical model where the normal distribution for z_{2t} in (15) is replaced by the t on 3.5 degrees of freedom. These estimates are in the supplemental material for this article at the *JASA* website. The posterior distribution of $\boldsymbol{\theta}$ is not affected in any substantive way. If anything, the evidence against the habit model is slightly stronger because the already overly large risk premium of $11.14\% - 1.21\% = 9.93\%$ in Table 1 increases to $11.07\% - 0.73\% = 10.34\%$ and the already overly large mean of

consumption growth of 2.087% becomes 2.26%.

6 Conclusion

We considered a consumption based asset pricing model that uses habit persistence to over-come the known statistical inadequacies of the classical consumption based asset pricing model. We found that the habit model fits reasonably well and agrees with results reported in the literature if conditional heteroskedasticity is suppressed but that it does not fit nor do results agree if conditional heteroskedasticity is allowed to manifest itself. We also found that it is the preference parameters of the model that are most affected by the presence or absence of conditional heteroskedasticity, especially the risk aversion parameter.

To obtain these results we proposed and implemented a general purpose Bayesian method-ology for the analysis of complex models from the sciences. It relies on the ability to simulate from the scientific model, upon the availability of substantive prior information, and upon the willingness to use that prior information. Analysis is carried out by means of a richly parameterized statistical model $f(\cdot|\cdot, \boldsymbol{\eta})$ that is viewed as being the correct description of the distribution of the data. To assess the scientific model, we view it as imposing a severe prior $\pi_\kappa(\boldsymbol{\eta})$ on the statistical model. The correctness of the scientific model is assessed by relaxing $\pi_\kappa(\boldsymbol{\eta})$ and assessing the posteriors of scientifically meaningful functionals of $f(\cdot|\cdot, \boldsymbol{\eta})$. If lo-cation does change by a scientifically meaningful amount as $\pi_\kappa(\boldsymbol{\eta})$ is relaxed, then the model is supported. If scale increases as $\pi_\kappa(\boldsymbol{\eta})$ is relaxed, then the model has empirical content.

References

Bansal, R., Gallant, A. R., and Tauchen, G., (2007), "Rational Pessimism, Rational Exu-berance,and Asset Pricing Models," *Review of Economic Studies*, 5, 523–590.

Campbell, J. Y., (2003), "Consumption Based Asset Pricing," in *Handbook of the Economics of Finance*, Vol. IB, eds. G. Constantinides, M. Harris, and R. Stulz, Amsterdam: North-Holland, 803-887.

Campbell, J. Y., and Cochrane, J., (1999), "By Force of Habit: A Consumption-based Explanation of Aggregate Stock Market Behavior," *Journal of Political Economy*, 107, 205–251.

Dejong, D. N., Ingram, B. F., and Whiteman, C. H., (1996), "A Bayesian Approach to Calibration," *Journal of Business and Economic Statistics*, 14, 1–9.

Dejong, D. N., Ingram, B. F., and Whiteman, C. H., (2000), "Keynesian Impulses versus Solow Residuals: Itentifying Sources of Business Cycle Fluctuations," *Journal of Applied Econometrics*, 15, 311–329.

Del Negro, M., and Schorfheide, F., (2004), "Priors from General Equilibrium Models for VARS,' *International Economic Review*, 45, 643–673.

Gallant, A. R., (1987), "Identification and Consistency in Seminonparametric Regression," in ed. Bewley, T. F., *Advances in Econometrics Fifth World Congress, Volume 1*, New York: Cambridge University Press, 145–170.

Chernov, M., Gallant, A. R., Ghysels, E., and Tauchen, G., (2003), "Alternative Models for Stock Price Dynamics," *Journal of Econometrics*, 116, 225–257 .

Gallant, A. R., and Tauchen, G., (1996), "Which Moments to Match?" *Econometric Theory*, 12, 657–681.

Gamerman, D., and Lopes, H. F., (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference (2nd Edition)*, Boca Raton, FL: Chapman & Hall.

Geweke, J., (2005), *Contemporary Bayesian Econometrics and Statistics* New York: Wiley.

Gourieroux, C., Monfort, A., and Renault, E., (1993) "Indirect Inference," *Journal of Applied Econometrics,*, 8, S85–S118.

Guvenen, F., (2005), "A Parsimonious Macroeconomic Model for Asset Pricing: Habit Formation or Cross-sectional Heterogeneity?" Manuscript, Department of Economics, University of Minnesota.

Lucus, R. E., Jr., (1978), "Asset Prices in and Exchange Economy, *Econometrica*, 46, 1429–1445.

Mishkin, F. S., (1981) "The Real Rate of Interest: An Empirical Investigation," *Carnegie-Rochester Conference Series on Public Policy, The Cost and Consequences of Inflation*, 15, 151–200.

Poirier, D. J., (1988), "Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics," *Journal of Economic Perspectives*, 2, 121–144.

Smith, A. A., (1993), "Estimating Nonlinear Time Series Models Using Simulated Vector Autoregressions," *The Journal of Applied Econometrics*, 8, S63–S84.

Table 1. Parameter Estimates, Normal Errors

Parameter	BGT Estimates		Bayes-VAR			Bayes-GARCH		
Monthly	Estimate	Std. Err.	Mode	Mean	Std. Dev.	Mode	Mean	Std. Dev.
g	0.002116	0.000250	0.001639	0.001739	0.000258	0.001803	0.001780	0.000684
ψ_{11}	0.006151	0.000896						
ψ_{22}	0.036503	0.007716						
ρ_s	0.971900	0.015449						
μ_{dc}	-3.3587	0.0380						
r_{11}			0.006753	0.007326	0.000627	0.007254	0.007417	0.001903
r_{12}			0.001350	0.001451	0.000403	0.001350	0.001336	0.001068
r_{22}			0.003125	0.008109	0.006205	0.003125	0.018852	0.034435
ϕ	0.9853	0.0026	0.9861	0.9857	0.0024	0.9804	0.9818	0.0095
δ	0.9939	0.0005	0.9955	0.9937	0.0030	0.9898	0.9907	0.0070
γ	0.8386	0.2462	0.5726	0.9463	0.6179	1.0744	1.1747	1.7638
Annualized	Estimate	Std. Err.	Mode	Mean	Std. Dev.	Mode	Mean	Std. Dev.
g	2.539	0.0866	1.9672	2.087	0.0895	2.164	2.136	0.2369
σ	2.1308		2.3857	2.5870	0.2202	2.5589	2.6106	0.2513
ρ	0.1650		0.1960	0.1943	0.0508	0.1830	0.1773	0.0507
σ_w	12.9118		1.0825	2.8090	2.1496	1.0825	6.5306	4.4984
ϕ	0.8372	0.0090	0.8450	0.8412	0.0084	0.7890	0.8023	0.0328
δ	0.9292	0.0018	0.9477	0.9269	0.0102	0.8845	0.8934	0.0244
γ	0.8386	0.2462	0.5726	0.9463	0.6179	1.0744	1.1747	1.7638
$\mathcal{E}(r^a_{dt})$	6.54		6.00	7.58	0.6930	11.14	10.45	0.5487
$SDev(r^a_{dt})$	16.9		20.49	21.48	1.5343	24.22	26.14	2.4735
$\mathcal{E}(r^a_{ft})$	1.07		1.20	1.16	0.1389	1.21	0.99	0.1451
	$\chi^2(5) = 7.11$ (0.21)		reps $= 800,000$ by 8			reps $= 800,000$ by 8		

Notes: The values shown as monthy are the actual parameters of the habit model when simulated at a monthly frequency. To aid interpretation the location, scale, and discount parameters $g, \sigma, \sigma_w, \delta$ have been re-expressed as annual rates in percentage terms. The autoregressive parameter ϕ is adjusted from a monthly to annual frequency; ρ, γ do not require adjustment. BGT estimates, from Bansal, Gallant, and Tauchen (2004), use data on the price dividend ratio and the consumption dividend ratio in addition to consumption growth and stock returns and impose $\mathcal{E}(r^a_{f,t}) = 0.89\%$ and cointegration among consumption, dividends, and price. Variance parameters relate as

$$\mathrm{Var}\begin{pmatrix} c_t - c_{t-1} \\ d_t - d_{t-1} \end{pmatrix} = \begin{pmatrix} \sigma^2 & \rho\,\sigma\sigma_w \\ \mathrm{sym} & \sigma_w^2 \end{pmatrix} = \begin{pmatrix} r_{11}^2 + r_{12}^2 & r_{12}r_{22} \\ \mathrm{sym} & r_{22}^2 \end{pmatrix} = \begin{pmatrix} \psi_{11}^2 & \psi_{11}^2 \\ \mathrm{sym} & \psi_{11}^2 + 2\psi_{22}^2(1-\rho_s)^{-1} \end{pmatrix}$$

where BGT parameters μ_{dc}, ρ_s, and ψ_{22} are the location, autoregressive, and scale prameters of the cointegration relation between c_t and d_t. In the data, the mean of r^a_d is 6.02% and the standard deviation is 19.29%; for consumption growth the values are 1.95% and 2.24%. The mode of $\theta = (g, \sigma, \rho, \sigma_w, \delta, \gamma)$ is the mode of the multivariate posterior. All standard deviations are from marginal posteriors.
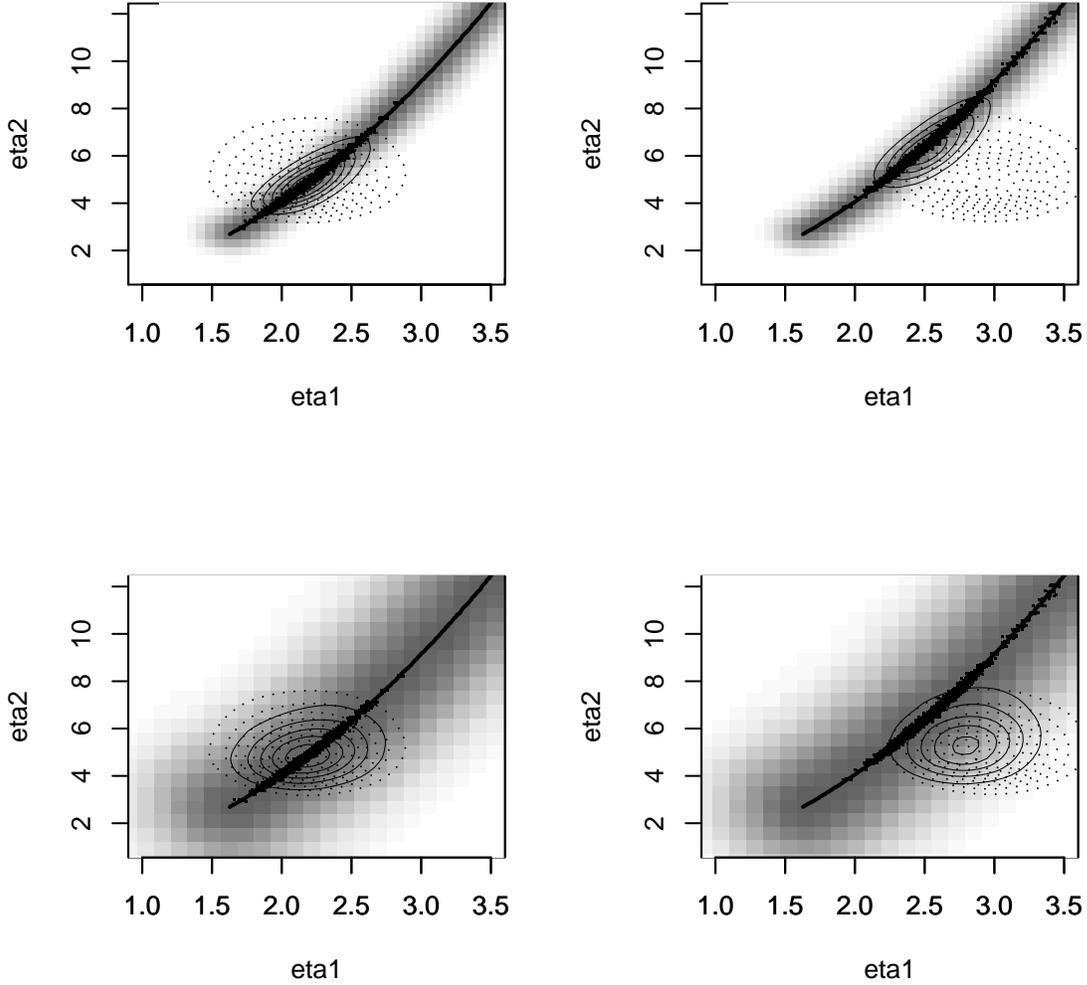
Figure 1. Priors and posteriors for the statistical model, tinker toy example.
The dotted lines are contours of the likelihood of the statistical model $f(y|x, \boldsymbol{\eta})$ of the tinker toy example. The line is the prior on $\boldsymbol{\eta}$ determined by the implied map $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta})$ from the parameters $\boldsymbol{\theta}$ of scientific model $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ to the parameters $\boldsymbol{\eta}$ of the statistical model. In the left panels the scientific model is true, in the right it is false. The thickness of the line is proportional to the posterior of $\boldsymbol{\eta}$. The prior $\pi(\boldsymbol{\eta})$ can be relaxed as indicated by the shading. The lower panels are more relaxed than the upper. The solid contours show the posterior under the relaxed prior. Relaxation causes the contours to enlarge in all cases. When the scientific model is false, the posterior shifts in search of the likelihood.

30
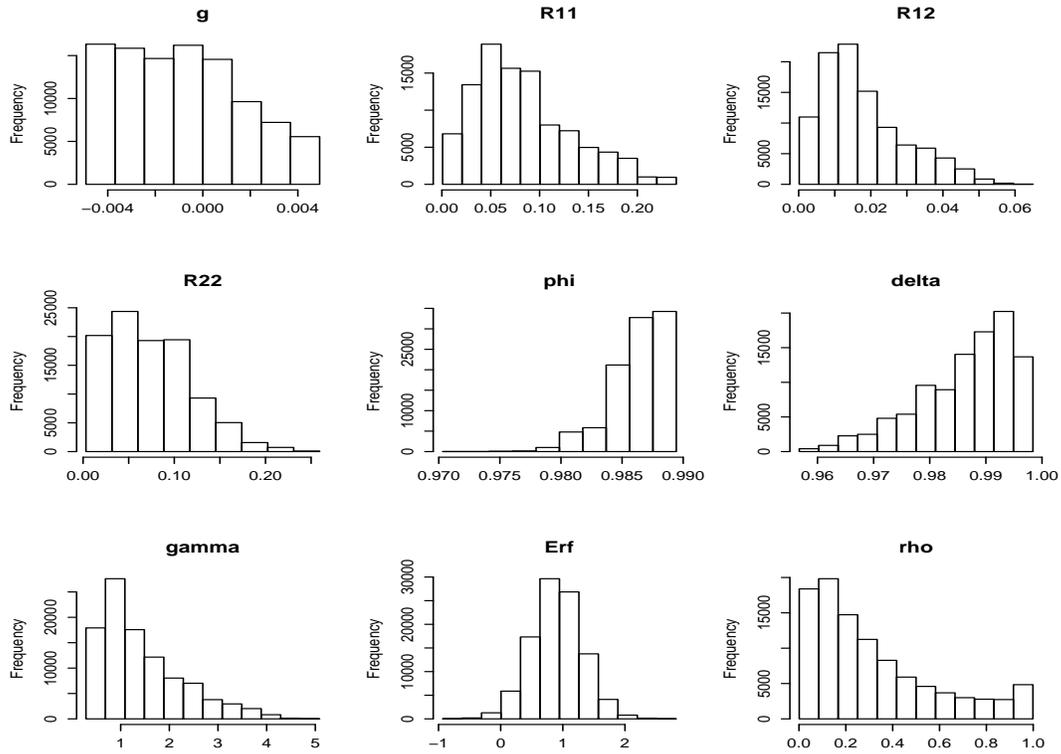
Figure 2. Priors distribution for the habit model example. Histograms of draws from the marginal priors of each of the seven elements of $\boldsymbol{\theta}$ followed by histograms for $\psi = \mathcal{E}(r_{ft}^a)$, and ρ.
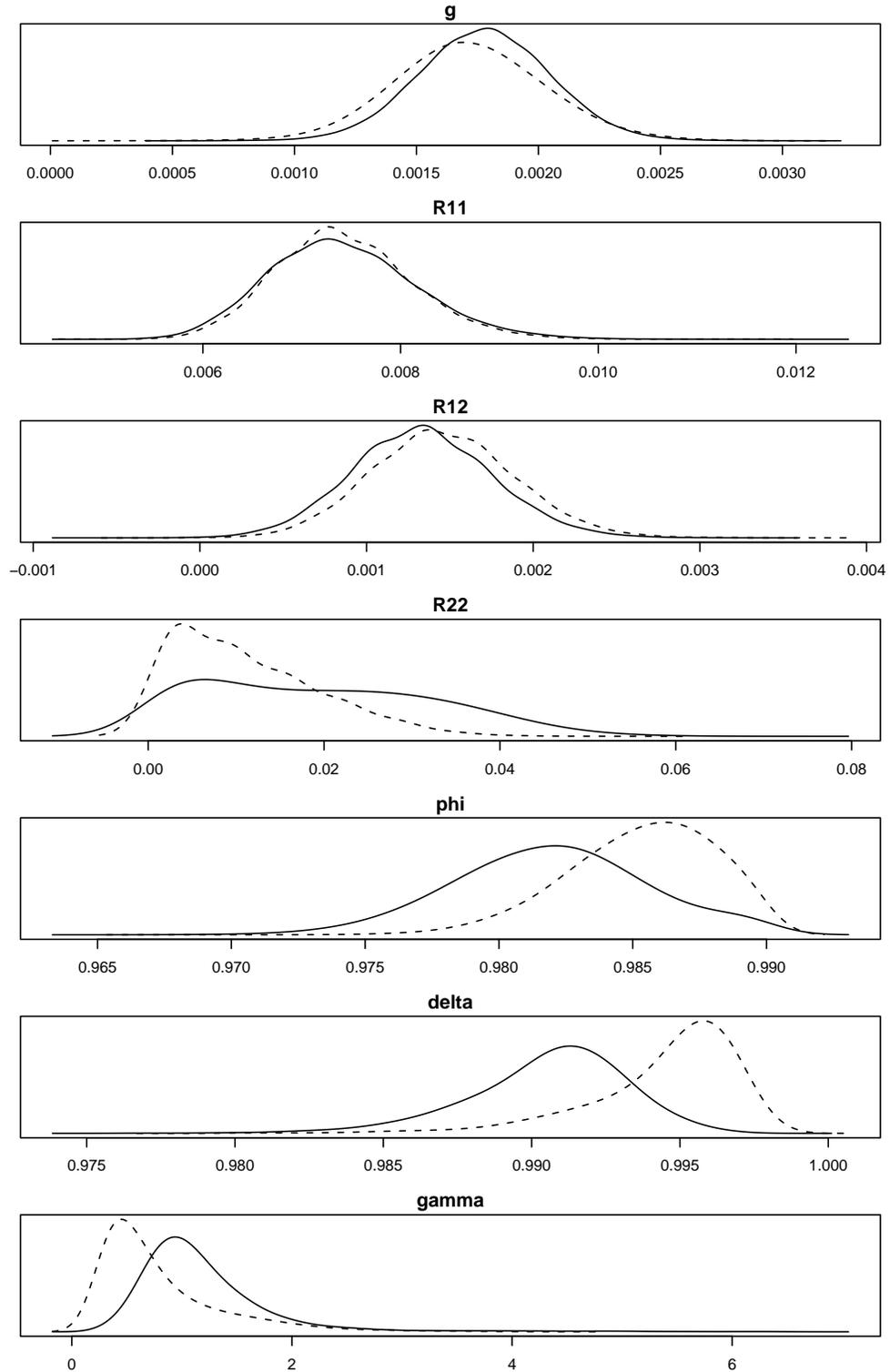
Figure 3. Density of the MCMC chain for $\boldsymbol{\theta}$. Shown is a kernel density estimate from iterates 1 to 800,000 by 8 of the MCMC chain for $\boldsymbol{\theta} = (g, r_{11}, r_{12}, r_{22}, \psi, \delta, \gamma)$ at the monthly frequency. The dashed line is for the Bayes-VAR chain and the solid line is for the Bayes-GARCH chain.
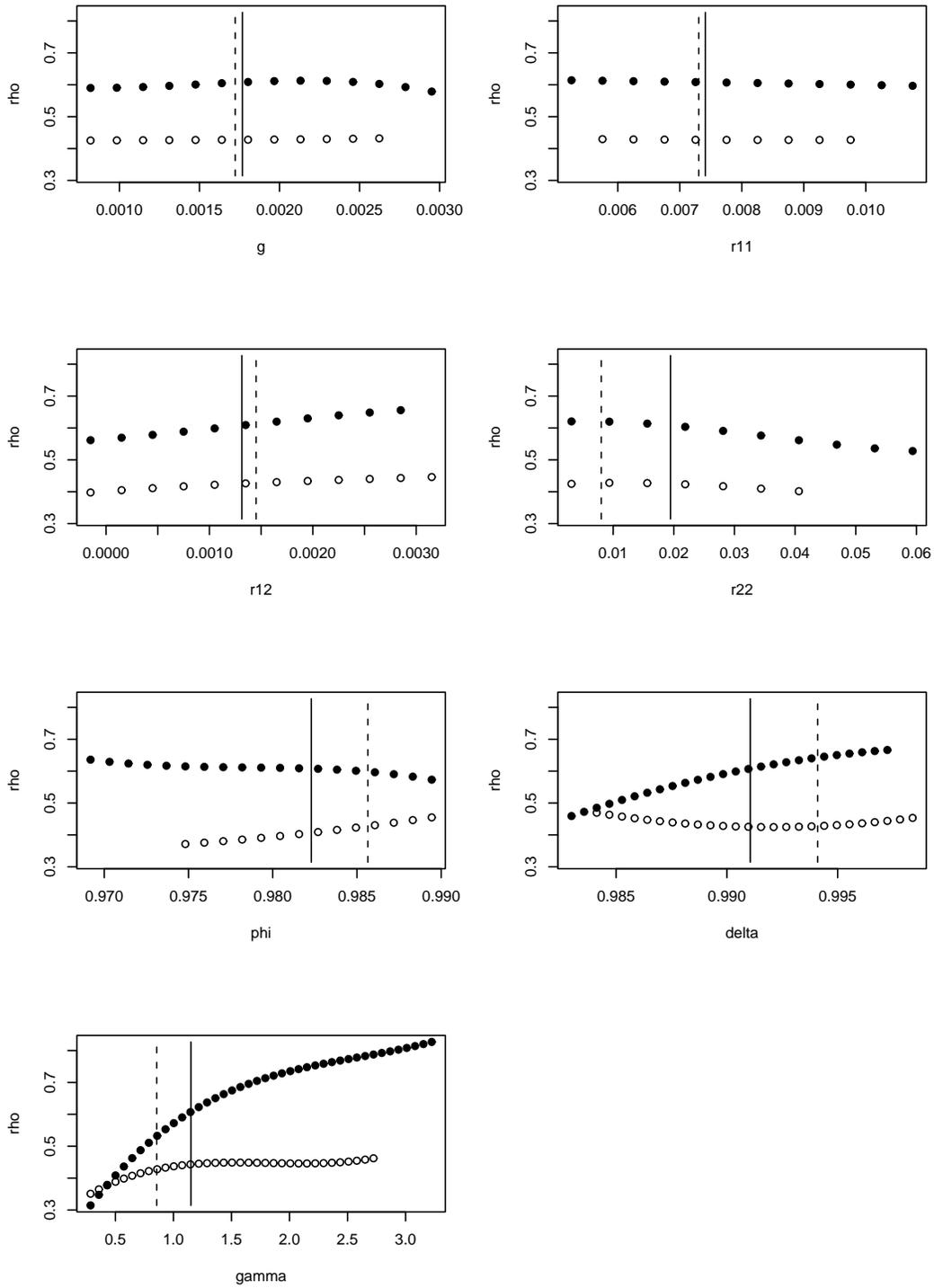
Figure 4. Conditional Correlation. The conditional correlation between annual consumption growth and stock returns for 2002 plotted against scientific model parameters $\boldsymbol{\theta} = (g, r_{11}, r_{21}, r_{22}, \phi, \delta, \gamma)$ as determined from the map $\boldsymbol{g} : \theta \mapsto \boldsymbol{\eta}$. Dots are for the GARCH statistical model; circles for the VAR. The solid vertical line is the GARCH posterior mean and the dashed VAR.
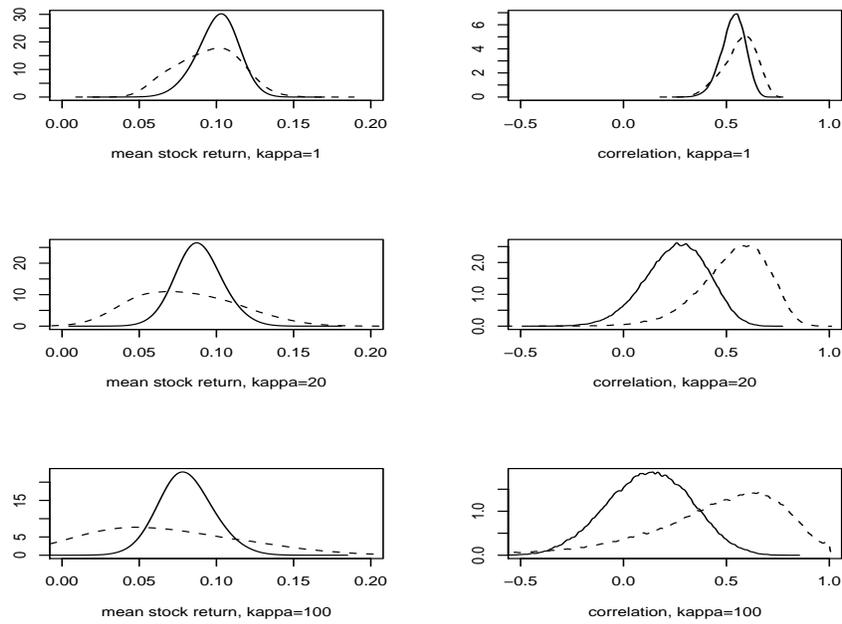
Figure 5. Priors and posterior for two functionals of the statistical model, habit model example. The statistical model is a bivariate GARCH model on annual consumption growth and stock returns. The scientific model is a habit persistence asset pricing model. The data covers 1929–2001. The left panels are the mean stock return for the period 2002–2102 implied by the GARCH model. The right panels are the conditional correlation between stock returns and consumption growth implied by the GARCH model for the year 2002. The prior is more relaxed in the lower panels than it is in the upper panels. The solid line is the posterior and the dashed line is the prior.