Reflections on the Probability Space Induced by Moment Conditions with Implications for Bayesian Inference*

A. Ronald Gallant
Penn State University

First draft: July 1, 2014
This draft: April 22, 2015
Forthcoming: *Journal of Financial Econometrics*
Available at www.aronaldg.org/papers/reflect.pdf

*Address correspondence to A. Ronald Gallant, P.O. Box 659, Chapel Hill NC 27514, USA, phone 919-428-1130; email aronldg@gmail.com.

Abstract

Often a structural model implies that certain moment functions expressed in terms of data and model parameters follow a distribution. An assertion that moment functions follow a distribution logically implies a distribution on the arguments of the moment functions. This fact would appear to permit Bayesian inference on model parameters. The classic example is an assertion that the sample mean centered at a parameter and scaled by its standard error has Student's t-distribution followed by an assertion that the sample mean plus and minus a critical value times the standard error is a Bayesian credibility interval for the parameter. This paper studies the logic of such assertions. The main finding is that if the moment functions have one of the properties of a pivotal, then the assertion of a distribution on moment functions coupled with a proper prior does permit Bayesian inference. Without the semi-pivotal condition, the assertion of a distribution for moment functions either partially or completely specifies the prior. In this case Bayesian inference may or may not be practicable depending on how much of the distribution of the constituents remains indeterminate after imposition of a non-contradictory prior. An asset pricing example that uses data from the US economy illustrates the ideas.

Keywords and Phrases: Moment functions, Structural Models, Bayesian inference

JEL Classification: C32, C36, E27

1 Introduction

The idea that a moment function can be used to make probability statements on its constituent random variables is at least as old as Fisher (1930). Fisher's assertion can be paraphrased in the context of this paper as follows: For observed data $x = (x_1, ..., x_n)$, if

$$t = \frac{\bar{x} - \theta}{s/\sqrt{n}},$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}),$$

has Student's t-distribution on $n-1$ degrees freedom, then

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \theta < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{1}$$

is a valid $(1-\alpha)\times 100\%$ credibility interval for θ, where $t_{\alpha/2}$ denotes the $1-\alpha/2$ quantile of Student's t-distribution on $n-1$ degrees freedom. The thinking underlying this construction is that an assumption of a distribution for t induces a joint distribution on the constituent random variables $(x_1, ..., x_n, \theta)$. From the joint one can obtain the conditional for θ given $(x_1, ..., x_n)$ and thereby make conditional probability statements on θ. Whether or not (1) could be regarded as valid Bayesian credibility interval without the need to specify a prior and whether or not this was what Fisher meant to say was controversial in its day.

It is immediately obvious that (1) cannot be a Bayesian credibility interval when the situation is as just stated because intervals of the form $a < \theta < b$ are not preimages of t and therefore cannot be assigned probability using only the assertion that t has the Student's t-distribution on $n-1$ degrees freedom. More is required. Our analysis in Section 3 implies that imposing a proper prior will suffice. The reason that imposing a prior is a remedy is that it has effect of enlarging the collection of sets to which probability can be assigned to include the intervals $a < \theta < b$.

This paper is a general consideration of an expanded view of the above situation: If one specifies a set of moment functions collected together into a vector $\bar{m}(x, \theta)$ of dimension M, where x is a statistic of dimension K, regards θ of dimension p as random, and asserts that some transformation $Z(x, \theta)$ of them has distribution $\Psi(z)$, then what is required to use

this information and possibly a prior to make valid Bayesian inferences? It is clear what the assertion has done: It has induced a probability measure on the preimages of Z. The answer to the question just posed depends on whether or not this probability space induces a reasonable notion of a likelihood for x given θ. This we investigate in Section 3.

In some instances one does not have to reflect on these issues. For example, if x is a sufficient statistic and a structural model clearly implies a conditional distribution for x given θ, then one immediately has a likelihood and can proceed directly to Bayesian inference. We are concerned with situations where the structural model does not imply exogeneity of θ, or one prefers not to rely on an assumption of exogeneity, or one cannot construct a likelihood at all due to the complexity of the model, or one does not trust the numerical approximations needed to construct a likelihood.

An example where the structural model implies that the straightforward conditional approach of the previous paragraph is not logically correct is an asset pricing model that states that the price of an asset is the conditional expectation of the future payoff to the asset times a stochastic discount factor where the conditioning is on information currently available to the investor. Suppose, within the Bayesian paradigm, one wishes to estimate the value of the stochastic discount factor at a point in time using gross returns $R_i = (P_{i,t} + D_{i,t})/P_{i,t-1}$ on several assets, $i = 1, \ldots, n$, where $P_{i,t}$ is price at time t and $D_{i,t}$ is the dividend paid since time $t-1$. If θ_t denotes the stochastic discount factor at time t and x the observed gross returns $R_{i,t}$, then the moment function

$$\bar{m}(x, \theta_t) = 1 - \frac{1}{n} \sum_{i=1}^{n} \theta_t R_{i,t}$$

has time $t-1$ conditional expectation zero, hence unconditional expectation zero, and one might expect $\sqrt{n}\, \bar{m}(x, \theta_t)$ to be approximately normally distributed. For Bayesian inference one needs the conditional density of x given θ_t in order to have a likelihood. However, θ_t is an endogenous random variable so that inferring a conditional distribution $p(x \,|\, \theta_t)$ solely from the situation as just stated takes some thought. As to a prior, asset pricing models often assume that θ_t is a function of its own past, e.g., $p(\theta_t \,|\, \theta_{t-1})$. If one can infer a likelihood $p(x \,|\, \theta_t)$, then one can use something such as $p(x \,|\, \theta_t) p(\theta_t \,|\, \theta_{t-1}) p(\theta_{t-1})$ as the basis for Bayesian inference. This formulation of the prior has introduced a second parameter θ_{t-1}

with prior $p(\theta_{t-1})$ that needs to be estimated.

2 Applications

As is usually the case, applications precede theory. An instance is Duan and Mela (2009). The typical implementation of Bayesian inference using method of moments proceeds as follows.

One has vector-valued observations x_t, $t = 1, 2, \ldots, n$, and has a parameter vector θ to be estimated. Let x denote the data arranged as a matrix with columns x_t. One sets forth moment functions $m(x_t, \theta)$ of dimension M and computes their mean

$$\bar{m}(x, \theta) = \frac{1}{n} \sum_{t=1}^{n} m(x_t, \theta).$$

The structural model implies that at the true value θ^o the unconditional expectation of the mean is zero, i.e., $\mathcal{E}\bar{m}(x, \theta^o) = 0$, and that θ^o is the only value of θ for which this is true. Put

$$Z(x, \theta) = \sqrt{n} \left[W(x, \theta)\right]^{-\frac{1}{2}} \left[\bar{m}(x, \theta)\right], \tag{2}$$

where

$$W(x, \theta) = \frac{1}{n} \sum_{t=1}^{n} \left[m(x_t, \theta) - \bar{m}(x, \theta)\right] \left[m(x_t, \theta) - \bar{m}(x, \theta)\right]' \tag{3}$$

and $[W(x, \theta)]^{-\frac{1}{2}}$ denotes the inverse of the Cholesky factorization of $W(x, \theta)$. If the $m(x_t, \theta)$ are serially correlated one will have to use a HAC (heteroskedatic, autoregressive invariant) variance matrix estimate (Gallant, 1987, p. 446, 533) instead. In this case it is essential that residuals $e_t = m(x_t, \theta) - \bar{m}(x, \theta)$ be used to form the estimate as in (3) rather than relying on $\mathcal{E}\bar{m}(x, \theta) = 0$, which only holds at the true value $\theta = \theta^o$.

Then one asserts that

$$p(x \,|\, \theta) = (2\pi)^{-\frac{M}{2}} \exp\left\{-\frac{n}{2} \bar{m}'(x, \theta) [W(x, \theta)]^{-1} \bar{m}(x, \theta)\right\}, \tag{4}$$

is a likelihood and proceeds directly to Bayesian inference using a prior $p^*(\theta)$.

The assertion (4) amounts to a belief that $Z(x, \theta)$ is normally distributed. This is not essential, one could assume that $Z(x, \theta)$ has a multivariate Student-t distribution or some other plausible distribution. Or, one could use some $Z(x, \theta)$ other than (2).

The usual computational method is MCMC (Markov chain Monte Carlo) for which the best known reference in econometrics is Chernozhukov and Hong (2003). A more comprehensive reference is Gamerman and Lopes (2006). Chernozhukov and Hong term (4) the Laplace density. MCMC generates a (correlated) sample from the posterior. From this sample one can compute the posterior mean and standard deviations, which are the usual statistics used to report Bayesian results. One can also compute the marginal likelihood from the chain (Newton and Raftery (1994)), which is used for Bayesian model comparison.

The estimator for frequentist inference is

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ \bar{m}'(x,\theta)[W(x,\theta)]^{-1}\bar{m}(x,\theta) \right\},$$

often, but not always, with θ that appears in $W(x,\theta)$ set to some preliminary estimate. One method for computing $\hat{\theta}$ is to apply MCMC to (4) with $p^*(\theta) \equiv 1$ (Chernozhukov and Hong (2003)).

The methods set forth in this section are used to estimate the habit persistence asset pricing model (Campbell and Cochrane (1999)) from US data in Section 4. We next examine the plausibility of the assertion that (4) can be regarded as a likelihood for the purpose of Bayesian inference.

3 The Likelihood Induced by Moment Functions

In what follows, prior probability is represented by a random variable Λ that has realization θ that lies in a parameter space Θ. Similarly for data, X is the random variable with realization x that lies in a sample space $\mathcal{X}$. As to dimensions, the parameter space Θ is a subset of $\mathbb{R}^p$ and the sample space $\mathcal{X}$ is a subset of $\mathbb{R}^K$.

The conceptual framework is that X and Λ are jointly distributed on a probability space $(\mathcal{X} \times \Theta, \mathcal{C}^o, P^o)$ determined by a structural model and a prior, where $\mathcal{C}^o$ denotes the Borel subsets of $\mathbb{R}^{K+p}$ intersected with $\mathcal{X} \times \Theta$. In simple cases the structural model determines a likelihood, $p^o(x \,|\, \Lambda = \theta)$, and the likelihood times the density function of the prior, $p^*(\theta)$, determines P^o. For compatibility with later notions we use the notation $p^*(\theta)$ for the prior although it is also the marginal density for Λ under P^o and could therefore logically be

denoted by $p^o(\theta)$. We presume the existence of $(\mathcal{X} \times \Theta, \mathcal{C}^o, P^o)$ but do not assume that it has the simple structure just described or, if it does, that the simple structure is known.

A common situation that requires consideration of the notions that follow is that deriving the likelihood from a structural model is analytically intractable and one cannot verify that the numerical approximations one would have to make to circumvent the intractability are sufficiently accurate. Therefore one uses features of the structural model, such as first order conditions, to derive moment functions $\bar{m} : (x, \theta) \mapsto \mathbb{R}^M$. A transformation of the moment functions, such as $Z(x, \theta) = \sqrt{n}\,[W(x, \theta)]^{-1/2}\bar{m}(x, \theta)$, is a random variable defined on $(\mathcal{X} \times \Theta, \mathcal{C}^o, P^o)$. We presume that the distribution of the transformed moment functions can be plausibly asserted to be $\Psi(z)$. Typically, verification that $\mathcal{E}^o[\bar{m}(X, \theta)] = 0$ has a unique solution when viewed as a function of θ is critical to the plausibility of the assertion that the distribution of Z is Ψ. The assertion that the distribution of Z is Ψ can be used to derive a probability space $(\mathcal{X} \times \Theta, \mathcal{C}^*, P^*)$ such that $P^*(C) = P^o(C)$ for $C \in \mathcal{C}^*$. The probability space $(\mathcal{X} \times \Theta, \mathcal{C}^*, P^*)$ can be used as a substitute for $(\mathcal{X} \times \Theta, \mathcal{C}^o, P^o)$ for the purpose of Bayesian inference. In this section we derive $(\mathcal{X} \times \Theta, \mathcal{C}^*, P^*)$ given Z and Ψ from first principles.

This section is similar to Section 3 of Gallant, Giacomini, and Ragusa (2014). The development there was to derive a measurement density for a state space model from moment functions for use in particle filtering. While the pairing of a measurement density and a transition density of a state space model and the pairing of a likelihood and a prior of a Bayesian model have similarities, there is enough difference to make it necessary to redevelop the ideas in a Bayesian context. The development proceeds first by discrete examples to avoid measurablility issues then to the general case. The extension to the general case is straightforward once one has seen the examples. An early development of these ideas appeared in Gallant and Hong (2007) but application to the present context is not straightforward because their model had a hierarchical structure, like a state space model, and they glossed over some essential details.

This section makes the following points.

- The transformed moment functions induce a probability measure P on a σ-algebra $\mathcal{C}$ containing sets C that have elements (x, θ). $\mathcal{C}$ is the smallest σ-algebra that contains

the preimages of Z. Typically $\mathcal{C}$ is coarse in the sense that it does not contain all the Borel sets. In particular, if $\mathcal{C}$ does not contain the rectangles $R_B = (\mathbb{R}^K \times B) \cap (\mathcal{X} \times \Theta)$, where B is a Borel subset of $\mathbb{R}^p$, then the probability space $(\mathcal{X} \times \Theta, \mathcal{C}, P)$ cannot be used for Bayesian inference.

- Specification of a prior for θ allows one to embed $\mathcal{C}$ within a σ-algebra $\mathcal{C}^*$ that contains the rectangles R_B and to define a probability measure P^* on $\mathcal{C}^*$ that agrees with both P and P^o on $\mathcal{C}$. The probability space $(\mathcal{X} \times \Theta, \mathcal{C}^*, P^*)$ can be used for Bayesian inference.

- Complications arise in applications if the transformation of the moment functions does not have some of the properties of a pivotal.

3.1 A Probability Distribution Induced by Moment functions

Consider the assertion that the probability distribution of a moment function D is as shown in Table 1. In Table 1, D models the difference $D = X - \Lambda$ between the toss of two correlated, six-sided dice X and Λ. The expectation of D is zero. One wishes to determine the posterior for Λ; i.e., the conditional distribution Λ given X. The first step toward this goal is to determine the likelihood, which is the conditional distribution of X given Λ.

(Table 1 about here)

The sets C_d shown in Table 1 are mutually exclusive and exhaustive. They partition the domain of D as finely as possible by preimages of D because they are the preimages of the singleton sets from the range of D. For a probability space with this structure, one conditions on knowing that the random variable Λ has the value θ by conditioning on the union of all preimages C_d that contain the point (x, θ) for some x. Denote this union by O_θ. O_θ is the union of all preimages of singleton sets that can occur if $\Lambda = \theta$ is known to have occurred. For the specific case shown in Table 1, the conditional probability density is

$$P(D = d \,|\, \Lambda = \theta) = \frac{P(C_d \cap O_\theta)}{P(O_\theta)}, \tag{5}$$

where C_d is the preimage of d under D, as displayed in Table 1. Let $\mathcal{C}$ be the smallest σ-algebra that contains the preimages $\{C_d : d = -5, \ldots, 5\}$. In this instance, $\mathcal{C}$ consists of the empty set $\emptyset$ and all possible unions of the sets C_d.

One is accustomed to the case where O_θ is the rectangle $\mathcal{X} \times \{\theta\}$, which in this example would be $R_\theta = \mathbb{D} \times \{\theta\}$ with $\mathbb{D} = \{1, 2, 3, 4, 5, 6\}$. But, in this example, $\mathcal{C}$ does not include the rectangles R_θ. If the σ-algebra over which probability is defined does not contain all the rectangles then O_θ need not take the form $\mathbb{D} \times \{\theta\}$. Nonetheless, the principle expressed in (5) remains valid.

An alternate expression for $P(D = d \,|\, \Lambda = \theta)$, useful below, is

$$P(D = d \,|\, \Lambda = \theta) = \frac{\sum_{x=1}^{6} I_{C_d}(x, \theta) P(D = d)}{\sum_{d=-5}^{5} \sum_{x=1}^{6} I_{C_d}(x, \theta) P(D = d)}. \tag{6}$$

This expression relies on the fact that for this example x and d are in a one-to-one correspondence once θ is fixed. This construction has also induced a "marginal" distribution

$$Q(\Lambda = \theta) = P(O_\theta) = \sum_{d=-5}^{5} \sum_{x=1}^{6} I_{C_d}(x, \theta) P(D = d). \tag{7}$$

The sense in which (7) defines a marginal is

$$P(D = d) = \sum_{\theta=1}^{6} P(D = d \,|\, \Lambda = \theta) Q(\Lambda = \theta).$$

A notion of marginal on Λ can be regarded as a partial specification of a prior. We will explore this issue in the example in Subsection 3.6 where the moment functions induces both a likelihood and a prior.

Any $\mathcal{C}$-measurable f must be constant on the preimages. For such f the formula

$$\mathcal{E}(f \,|\, \Lambda = \theta) = \sum_{x=1}^{6} f(x, \theta) \sum_{d=-5}^{5} I_{C_d}(x, \theta) P(D = d \,|\, \Lambda = \theta) \tag{8}$$

can be used to compute conditional expectation because f can be regarded as a function of d and the right hand side of (8) equals

$$\sum_{d=-5}^{5} f(d) P(D = d \,|\, \Lambda = \theta).$$

Equation (8) implies that we can view $P(D = d)$ as defining a conditional density function

$$P(X = x \,|\, \Lambda = \theta) = \sum_{d=-5}^{5} I_{C_d}(x, \theta) P(D = d \,|\, \Lambda = \theta) \tag{9}$$

that is a function of x as long as we only use it in connection with $\mathcal{C}$-measurable f.

Intuitively what is going on here is that if θ is held constant, then X becomes a transform of D, which has a known conditional density. This can be seen more easily by rewriting (9) as

$$P(X = x \,|\, \Lambda = \theta) = \frac{P(D = x - \theta)}{\sum_{x=1}^{6} P(D = x - \theta)}. \tag{10}$$

Note also that $Q(\Lambda = \theta) = \sum_{x=1}^{6} P(D = x - \theta)$.

Similar considerations define $P(D = d \,|\, X = x)$,

$$P(\Lambda = \theta \,|\, X = x) = \frac{P(D = x - \theta)}{\sum_{\theta=1}^{6} P(D = x - \theta)}.$$

and $Q(X = x) = \sum_{\theta=1}^{6} P(D = x - \theta)$.

$P(\Lambda = \theta \,|\, X = x)$ is not a useful posterior density because the rectangles R_θ are not in $\mathcal{C}$, which means that one cannot make inferences regarding θ. For example, one cannot compute posterior probability for $\{\theta \,:\, a < \theta \leq b\}$ and therefore cannot determine a credibility interval.

On the other hand, were it the case that $O_\theta = \mathcal{X} \times \{\theta\}$, then we would be done, because the sets $\mathcal{X} \times B$, where $B \subset \mathbb{D}$, would be in $\mathcal{C}$. That is, if $O_\theta = \mathcal{X} \times \{\theta\}$, which is a property of the preimages of the transformed moment functions and not of the distribution that one asserts for them, then an assertion that a set of transformed moment functions follows a particular distribution implies a likelihood, a prior, and a posterior directly. No further input is necessary. We shall see an example of this in Subsection 3.6.

The ideal situation would be when a moment function specification determines a likelihood but not the prior, leaving one free to conduct Bayesian inference in the traditional fashion. It is the images of $\mathcal{X} \times \{\theta\}$ that determines when this situation occurs as we shall see in Subsection 3.4.

3.2 Dominating Measure

With respect to Table 1, consider the situation where X is itself a moment $X = X_1 + X_2$, where the range of both X_1 and X_2 are the integers. Let,

$$B_s = \{(x_1, x_2) \,:\, x_1 + x_2 = s;\, x_1, x_2 = 0, \pm 1, \pm 2, \ldots\}$$

for $s = 1, 2, \ldots, 6$. Then the preimages C_d listed in Table 1 become, instead,

$$\begin{aligned} C'_{-5} &= \{(x_1, x_2, 6) \,:\, (x_1, x_2) \in B_1\} \\ C'_{-4} &= \{(x_1, x_2, 5) \,:\, (x_1, x_2) \in B_1\} \cup \{(x_1, x_2, 6) \,:\, (x_1, x_2) \in B_2\} \\ &\vdots \end{aligned}$$

The difficulty we run into is that we do not have an obvious dominating measure with which to integrate the conditional density $P[(X_1, X_2) = (x_1, x_2) \,|\, \Lambda = \theta]$. One way to circumvent the difficulty is as follows. Given θ, for each $s = 1, 2, \ldots, 6$, choose a representer $(x_1, x_2)^* \in B_s$ to label B_s. The dominating measure puts mass one on these six representers and mass zero on all other pairs of integers.

For our purposes we can gloss over the issue of a dominating measure. We never use it so that mere existence suffices. All that is required to construct a dominating measure is a labeling scheme that can put preimages of singleton sets into a one-to-one correspondence with a representative x when θ is given.

3.3 Introduction of a Prior

Continuing with the example of Table 1, we now assign prior probability $P^*(R_\theta) = \frac{1}{6}$ to the rectangles $R_\theta = \mathbb{D} \times \{\theta\}$, where $\theta \in \mathbb{D} = \{1, 2, 3, 4, 5, 6\}$. Put $P^*(C) = P(C)$ for $C \in \mathcal{C}$. Let $\mathcal{C}^*$ denote the smallest σ-algebra that contains both $\{C_d\}_{d=-5}^{5}$ and $\{R_\theta\}_{\theta=1}^{6}$. In principle the definition of P^* can be extended to all sets in $\mathcal{C}^*$. Let $(\mathbb{D} \times \mathbb{D}, \mathcal{C}^*, P^*)$ denote the extended probability space. In this instance, the singleton sets $\{(x, \theta)\}$ are in $\mathcal{C}^*$ so that under P^* conditional probability has its conventional definition

$$\begin{aligned} P^*(X = x \,|\, \Lambda = \theta) &= \frac{P^*(\{(x, \theta)\})}{P^*(R_\theta)} \\ P^*(\Lambda = \theta \,|\, X = x) &= \frac{P^*(\{(x, \theta)\})}{P^*(R_x)}, \end{aligned}$$

where $R_\theta = \mathbb{D} \times \{\theta\}$; $\theta \in \mathbb{D}$ and $R_x = \{x\} \times \mathbb{D}$; $x \in \mathbb{D}$.

A difficulty is that the information in Table 1 and the knowledge that $P^*(R_\theta) = \frac{1}{6}$ is not enough to deduce $P^*(\{(x, \theta\})$ because that knowledge implies a singular system of nine

equations in sixteen unknowns

$$\begin{aligned}
\frac{4}{18} &= \sum_{i=1}^{5} P^*(\{(i,i+1)\}) \qquad (11)\\
\frac{10}{18} &= \sum_{i=1}^{6} P^*(\{(i,i)\})\\
\frac{4}{18} &= \sum_{i=1}^{5} P^*(\{(i+1,i)\})\\
\frac{1}{6} &= P^*(\{(1,1)\}) + P^*(\{(2,1)\})\\
\frac{1}{6} &= P^*(\{(1,2)\}) + P^*(\{(2,2)\}) + P^*(\{(3,2)\})\\
\frac{1}{6} &= P^*(\{(2,3)\}) + P^*(\{(3,3)\}) + P^*(\{(4,3)\})\\
\frac{1}{6} &= P^*(\{(3,4)\}) + P^*(\{(4,4)\}) + P^*(\{(5,4)\})\\
\frac{1}{6} &= P^*(\{(4,5)\}) + P^*(\{(5,5)\}) + P^*(\{(6,5)\})\\
\frac{1}{6} &= P^*(\{(5,6)\}) + P^*(\{(6,6)\})
\end{aligned}$$

after taking into account that Table 1 implies that only the sixteen $P^*(\{(x,\theta\})$ that appear in (11) can be non-zero. The sum of the first three equations in (11) equals the sum of the last six so there are effectively only eight equations in sixteen unknowns.

One way to resolve this difficulty is to estimate the probabilities in (11) along with θ by assigning prior probability $0 \leq P^*(\{(x,\theta)\}) \leq \frac{1}{6}$ to eight of the $P^*(\{(x,\theta)\})$ in (11), using (11) to solve for the remaining eight. This strategy actually works well in this instance as regards estimation of θ using MCMC when a subset of the probabilities are not identified and therefore posterior probabilities are determined by the prior. Which subset depends on the value of θ used to generate the data.

While the difficulty that a prior and an assertion of a distribution for the transformed moment functions may not completely determine $(\mathbb{D}\times\mathbb{D}, \mathcal{C}^*, P^*)$ can be circumvented in this instance, what is outlined above is not an attractive general strategy. When the number of undetermined probabilities is infinite in the discrete case or when Θ is a continuum, it is not clear how to proceed. The difficulty can be circumvented when $P(O_\theta) = 1$ as seen in the next example, which is actually a discretized variant of Fisher (1930).

3.4 An Example where $P(O_\theta) = 1$

Consider the case

$$\begin{aligned} P[Z(X,\Lambda) = z] &= \frac{1-p}{1+p}\, p^{|z|} \\ Z(X,\Lambda) &= X - \Lambda \\ X &\in \mathbb{N} \\ \Lambda &\in \mathbb{N} \\ \mathbb{N} &= \{\, 0,\, \pm 1,\, \pm 2,\, \ldots \,\} \end{aligned}$$

The preimages of $Z(x,\theta)$ are

$$C_z = \{(x,\theta) \,:\, x = z + \theta,\, \theta \in \mathbb{N}\} \qquad z \in \mathbb{N}$$

which lie on 45 degree lines in the (x,θ) plane. Given θ, for every $z \in \mathbb{N}$ there is an $x \in \mathbb{N}$ with $(x,\theta) \in C_z$ so every C_z can occur. Therefore $O_\theta = \cup_{z\in\mathbb{N}} C_z$ and $P(O_\theta) = 1$ for every $\theta \in \mathbb{N}$. Hence

$$P(Z = z \,|\, \Lambda = \theta) = \frac{P(C_z \cap O_\theta)}{P(O_\theta)} = P(C_z) = \frac{1-p}{1+p}\, p^{|z|}, \tag{12}$$

which does not depend on θ. Consequently,

$$P(X = x \,|\, \Lambda = \theta) = P(Z = x - \theta)$$

using logic analogous to that leading to equation (10).

This situation seems to be what occurs most often in applications because the chosen Z is often a pivotal or can be regarded as such in large samples. A pivotal has $P^o(O_\theta) = 1$ whence $P(O_\theta) = 1$. Much less than pivotal is actually required for $P(O_\theta) = 1$: If, for every $\theta \in \Theta$, the image of $\mathcal{X} \times \{\theta\}$ under $Z(\cdot,\theta)$ is the entire support of the distribution $\Psi(z)$ then $P(O_\theta) = 1$ for every $\theta \in \Theta$. Assumption 1 below formalizes the remarks in this paragraph.

When probability $P^*(R_\theta)$ is assigned to rectangles the extension of P to P^* is

$$\begin{aligned} P^*(X = x \,|\, \Lambda = \theta) &= P(Z = x - \theta) \\ P^*(X = x, \Lambda = \theta) &= P^*(X = x \,|\, \Lambda = \theta)\, P^*(R_\theta). \end{aligned} \tag{13}$$

The principal guiding the choice of solution (13) to equations analogous to (11) (not shown) is that the conditional probability of X given Λ should be the same under P^*_θ and P_θ. Similarly

for the conditional probability of Z given Λ. In the example of Subsection 3.3, the choice (13) was not available as a solution to (11) because equality of the conditional probability of Z given Λ under both P^*_θ and P_θ would be violated.

We next verify that the requisite conditions on P^*_θ are satisfied. Agreement on $\mathcal{C}$ is satisfied, i.e., $(\mathbb{N}\times\mathbb{N},\mathcal{C},P^*)=(\mathbb{N}\times\mathbb{N},\mathcal{C},P)$, because

$$P^*(Z=z)=\sum_{\theta\in\mathbb{N}}P^*(X=z+\theta,\Lambda=\theta)=P(Z=z)\sum_{\theta\in\mathbb{N}}P^*(R_\theta)=P(Z=z). \tag{14}$$

The correct probability is assigned to rectangles because

$$\sum_{x\in\mathbb{N}}P^*(X=x,\Lambda=\theta)=\sum_{x\in\mathbb{N}}P(Z=x-\theta)P^*(R_\theta)=P^*(R_\theta)\sum_{z\in\mathbb{N}}P(Z=z)=P^*(R_\theta).$$

Equations (14) and (12) imply that $P^*(Z=z\,|\,\Lambda=\theta)=P(Z=z\,|\,\Lambda=\theta)$.

3.5 The Abstraction

As mentioned above, we assume that X and Λ are jointly distributed on a probability space $(\mathcal{X}\times\Theta,\mathcal{C}^o,P^o)$ determined by a structural model and a prior. Because the structural model does not imply a likelihood in a straightforward manner, or because one does not trust the numerical approximations required to obtain a likelihood, or for whatever reason, we are in a situation where method of moments is an attractive strategy. To this end, we have specified an M-dimensional vector of transformed moment functions $Z(x,\theta)$ whose distribution is implied by the structural model and a prior. Denote this distribution by $\Psi(z)$, its density by $\psi(z)$, and its support by $\mathcal{Z}=\{z:\psi(z)>0\}$. The density function of the prior is denoted by $p^*(\theta)$.

Let $\mathcal{C}$ be the smallest σ-algebra containing the preimages $C=\{(x,\theta):Z(x,\theta)\in B\cap\mathcal{Z}\}$ where B ranges over the Borel subsets of $\mathbb{R}^M$. Because the distribution Ψ of $Z(X,\Lambda)$ is determined by the structural model and prior, the probability distribution P induced on $(\mathcal{X}\times\Theta,\mathcal{C})$ by Ψ can be presumed to satisfy $P(C)=P^o(C)$ for every $C\in\mathcal{C}$. Therefore, $(\mathcal{X}\times\Theta,\mathcal{C},P^o)=(\mathcal{X}\times\Theta,\mathcal{C},P)$, which implies that expectations $\mathcal{E}(f)$ are computed the same on either probability space for $\mathcal{C}$-measurable f.

We impose a requirement that provides Z with some of the properties of a pivotal:

ASSUMPTION 1 Let

$$C^{(\theta,z)} = \{x \in \mathcal{X} \,:\, Z(x,\theta) = z\}. \tag{15}$$

We assume that $C^{(\theta,z)}$ is not empty for any $(\theta, z) \in \Theta \times \mathcal{Z}$.

If $C^{(\theta,z)}$ is not empty, then for each $\theta \in \Theta$ and $z \in \mathcal{Z}$ we may choose a point $x^* \in \mathcal{X}$ for which

$$Z(x^*, \theta) = z.$$

The point x^* is the representer of $C^{(\theta,z)}$. Define

$$\Upsilon(z, \theta) = x^*. \tag{16}$$

Let x^o denote the observed realization of X and let $z^o = Z(x^o, \theta)$. For $z = z^o$ we shall choose the representer of $C^{(\theta,z)}$ to be x^o so that we have $x^o = \Upsilon[Z(x^o, \theta), \theta]$ for every $\theta \in \Theta$.

If $C^{(\theta,z)}$ is not empty, then every preimage of the form

$$C^z = \{(x, \theta) \,:\, Z(x,\theta) = z,\ x \in \mathcal{X},\ \theta \in \Theta\}$$

must contain $(\Upsilon(z,\theta), \theta)$. Thus, for every $z \in \mathcal{Z}$, C^z can occur if $\Lambda = \theta$ is known to have occurred. The sets C^z are in $\mathcal{C}$ and are a mutually exclusive and exhaustive partitioning of the preimage $Z^{-1}(\mathcal{Z})$ and no finer partitioning of $\mathcal{X} \times \Theta$ by sets from $\mathcal{C}$ is possible. Therefore the conditioning set for the event $\Lambda = \theta$ is

$$O_\theta = \cup_{z \in \mathcal{Z}} C^z = Z^{-1}(\mathcal{Z}),$$

which implies $P(O_\theta) = P^o(O_\theta) = \Psi(\mathcal{Z}) = 1$. ∎

Verification of Assumption 1 in an application is usually easy. For instance, if $Z(x, \theta)$ is given by (2) with (3) computed from residuals, then if one of the elements $x_{i,t}$ of x_t can assume any value in $\mathbb{R}$ and $m(x_t, \theta)$ is continuous in $x_{i,t}$ and is neither bounded from above nor below as $x_{i,t}$ varies, then Assumption 1 will be satisfied.

Let $\mathcal{C}^*$ be the smallest σ-algebra that contains all sets in $\mathcal{C}$ plus all rectangles of the form $R_B = (\mathbb{R}^K \times B) \cap (\mathcal{X} \times \Theta)$, where B is a Borel subset of $\mathbb{R}^p$. Motivated by the discussion in Subsection 3.4 we define a measure P^* on $\mathcal{C}^*$ by means of the densities

$$p^*(x \,|\, \Lambda = \theta) = \psi[Z(x, \theta)] \tag{17}$$

$$p^*(x,\theta) = p^*(x \mid \Lambda = \theta)\, p^*(\theta).$$

For given θ and $\mathcal{C}$-measurable f, which must be a function of the form $f[Z(x,\theta)]$, we define

$$\int f[Z(x,\theta)]\, p^*(x \mid \Lambda = \theta)\, dx = \int_{\mathcal{Z}} f(z)\, \psi(z)\, dz, \tag{18}$$

leaving the dominating measure dx on $\mathcal{X}^* = \{x^* : x^* = \Upsilon(z,\theta),\ z \in \mathcal{Z}\}$ unspecified. In particular, for $f(x,\theta) = I_B[Z(x,\theta)]$ where B is a Borel subset of $\mathbb{R}^M$, we have

$$\int I_B[Z(x,\theta)]\, p^*(x \mid \Lambda = \theta)\, dx = \int_{B\cap\mathcal{Z}} \psi(z)\, dz. \tag{19}$$

To each $C \in \mathcal{C}$ there is a Borel set B for which $C = \{(x,\theta) : Z(x,\theta) \in B\}$. Therefore,

$$P^*(C) = \int \left\{ \int I_B[Z(x,\theta)]\, p^*(x \mid \Lambda = \theta)\, dx \right\} p^*(\theta)\, d\theta = \int_{B\cap\mathcal{Z}} \psi(z)\, dz. \tag{20}$$

By construction,

$$P(C) = P^o(C) = \int_{B\cap\mathcal{Z}} \psi(z)\, dz.$$

For rectangles of the form $R_B = (\mathbb{R}^K \times B) \cap (\mathcal{X} \times \Theta)$, where B is a Borel subset of $\mathbb{R}^p$ we already have that

$$P^*(R_B) = \int_{B\cap\Theta} p^*(\theta)\, d\theta. \tag{21}$$

We conclude that P^*, P^o, and P assign the same values to $C \in \mathcal{C}$ and that P^* and P^o assign the same values to the rectangles R_B. For $C \in \mathcal{C}^*$ that cannot be computed using (20) and (21), define $P^*_\theta(C) = P^o_\theta(C)$. We cannot compute these additional probabilities but it does not matter because we never need to; their existence suffices. We now have

$$(\mathcal{X}\times\Theta, \mathcal{C}, P^o_\theta) = (\mathcal{X}\times\Theta, \mathcal{C}, P_\theta) = (\mathcal{X}\times\Theta, \mathcal{C}, P^*_\theta).$$

$$(\mathcal{X}\times\Theta, \mathcal{C}^*, P^o_\theta) = (\mathcal{X}\times\Theta, \mathcal{C}^*, P^*_\theta).$$

For any $\mathcal{C}$-measurable f, $\mathcal{E}(f)$ will be computed the same under any of these three probability measures: P^o_θ, P_θ, or P^*_θ. Similarly, for $\mathcal{C}^*$-measurable f, $\mathcal{E}(f)$ will be computed the same under P^o_θ, and P^*_θ.

3.6 An Example where the Moment Functions Induce a Posterior

Table 2 is the same example as Table 1 but with two moment functions. The two moment functions partition the sample space into singleton sets and therefore $\mathcal{C}$ contains all subsets of $\mathbb{D}^2$. The probabilities assigned to each $\{(x, \theta)\}$ displayed in Table 2 are a particular choice of a solution to equations (11). Because the probability space contains all subsets of $\mathbb{D}^2$, the probability space $(\mathbb{D}^2, \mathcal{C}, P)$, completely determines the likelihood and the prior.

Application of the formula

$$\begin{aligned} P(X = x \,|\, \Lambda = \theta) &= \sum_{d=-5}^{5} \sum_{e=-4}^{11} I_{C_{d,e}}(x, \theta) P(D, E = d, e \,|\, \Lambda = \theta) \\ &= \frac{P[(D, E) = (x - \theta, 2x - \theta)]}{\sum_{x=1}^{6} P[(D, E) = (x - \theta, 2x - \theta)]} \end{aligned}$$

gives the likelihood, which can be re-expressed as

$$P^*(X = 1 \,|\, \Lambda = 1) = P^*(X = 6 \,|\, \Lambda = 6) = 1 \quad \text{for } \theta = 1 \text{ or } 6$$

$$P^*(X = \theta - 1 \,|\, \Lambda = \theta) = P^*(X = \theta \,|\, \Lambda = \theta) = P^*(X = \theta + 1 \,|\, \Lambda = \theta) = \frac{1}{3} \quad \text{for } \theta \neq 1 \text{ or } 6.$$

$$P^*(X = x \,|\, \Lambda = \theta) = 0 \quad \text{otherwise.}$$

The prior is the marginal

$$P(\Lambda = \theta) = \frac{1}{6} \quad \text{for } \theta = 1, \ldots, 6$$

(Table 2 about here)

4 Illustration: The Habit Persistence Model

We illustrate the ideas with the habit persistence asset pricing model proposed by Campbell and Cochrane (1999) using US annual data over the period 1950–2013.

Throughout this section, lower case denotes the logarithm of an upper case quantity; e.g., $c_t = \log(C_t)$, where C_t is consumption during time period t, and $d_t = \log(D_t)$, where D_t is dividends paid during period t. The exceptions are the geometric return on an annual Treasury obligation $r_{ft} = \log(P_{f,t} + I_t) - \log P_{f,t-1}$ and the geometric stock return inclusive of dividends $r_{dt} = \log(P_{dt} + D_t) - \log P_{d,t-1}$, where $P_{f,t-1}$ is the price of an obligation at

the beginning of time period t that pays interest I_t at the end of period t, and P_{ft} its price at the end. This representation of bonds with terminal price P_{ft} different from 1 is an artifact of data construction and adjustment for inflation, all that is relevant is the ratio $(P_{dt} + I_t)/P_{d,t-1}$. Similarly, $P_{d,t-1}$ is the price of a stock at the beginning of time period t, and P_{dt} its price at the end. Means and standard deviations of the data are shown in Table 3

(Table 3 about here)

The driving processes for the habit persistence model are

$$\text{Consumption:} \quad c_t - c_{t-1} = g + v_t, \tag{22}$$

$$\text{Dividends:} \quad d_t - d_{t-1} = g + w_t,$$

$$\text{Random shocks:} \quad \begin{pmatrix} v_t \\ w_t \end{pmatrix} \sim \text{NID} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\sigma_w \\ \rho\sigma\sigma_w & \sigma_w^2 \end{pmatrix} \right].$$

The utility function is of the CRRA (constant relative risk aversion) style

$$\mathcal{E}_0 \left(\sum_{t=0}^{\infty} \delta^t \, \frac{(S_t C_t)^{1-\gamma} - 1}{1-\gamma} \right) \tag{23}$$

with corresponding intertemporal marginal rate of substitution

$$M_t = \delta \left(\frac{S_t C_t}{S_{t-1} C_{t-1}} \right)^{-\gamma}. \tag{24}$$

Habit persistence is implemented by two equations:

$$\text{Surplus ratio:} \quad s_t - \bar{s} = \phi \, (s_{t-1} - \bar{s}) + \lambda(s_{t-1}) v_{t-1}, \tag{25}$$

$$\text{Sensitivity function:} \quad \lambda(s) = \begin{cases} \frac{1}{\bar{s}} \sqrt{1 - 2(s - \bar{s})} - 1 & s \leq s_{\max} \\ 0 & s > s_{\max} \end{cases}. \tag{26}$$

Gross returns satisfy the equations

$$1 = \mathcal{E}_{t-1} \left[M_t (P_{d,t} + D_t) / P_{d,t-1} \right] \tag{27}$$

$$1 = \mathcal{E}_{t-1} \left[M_t (P_{f,t} + I_t) / P_{f,t-1} \right] \tag{28}$$

The parameter γ is a measure of curvature, which scales attitudes toward risk, and δ is the agent's discount factor. $\mathcal{E}_t$ is conditional expectation with respect to S_t, which is the state

variable; $s_t = \log(S_t)$. The quantities $\bar{S}$ and $s_{\max}$ can be computed from model parameters as $\bar{S} = \sigma\sqrt{\gamma/(1-\phi)}$ and $s_{\max} = \bar{s} + (1-\bar{S}^2)/2$. The variable $X_t = C_t(1-S_t)$ is called external habit. By substituting $S_t C_t = C_t - X_t$ in (23) one can see that utility is extremely low when consumption is close to X_t for $\gamma > 1$. Habit X_t is determined by past consumption as is seen by noting that $v_{t-1} = \log(C_{t-1}/C_{t-2}) - g$ in (22). Given the habit model's parameters

$$\theta = (g, \sigma, \sigma_w, \rho, \phi, \delta, \gamma), \tag{29}$$

one can compute all quantities above and, in particular, the realized values of v_t and w_t. The values calibrated by Campbell and Cochrane (1999) are

$$\theta^* = (0.0189, 0.015, 0.122, 0.2, 0.87, 0.89, 2.00). \tag{30}$$

The prior used here is

$$\pi(\theta) = \prod_{i=1}^{7} \mathrm{N}\left[\theta_i \,\middle|\, \theta_i^*,\ \left(\frac{\tau_i \theta_i^*}{1.96}\right)^2\right] \tag{31}$$

For, e.g., $\tau_i = 0.1$ the prior states that the marginal probability that θ_i is within 10% of θ_i^* is 95%. This prior is the same as the prior used in Aldrich and Gallant (2011) except that they pinned the risk free rate with a prior whereas we include treasury returns in the data. They used $\tau_i = 0.1$ save for ϕ and δ where they used $\tau_i = 0.001$. As documented by Aldrich and Gallant, ϕ and δ are not identified and must somehow be restricted when statistical methods are used to determine model parameters; for these two we find that any $\tau_i \leq 0.1$ suffices. We can use a larger value than Aldrich and Gallant because their estimation method required model solution which places more stringent requirements on admissible parameters values than does method of moments.

The habit model was admittedly reverse engineered to deal with the fact that consumption is too smooth to accurately price assets using CRRA utility and reasonable values of γ. This is fine as long as consumption remains smooth but in periods where it is not, i.e., the Great Depression and the Great Recession, the model not only cannot fit the data well but it also runs into numerical problems because economically implausible parameter values can become numerically plausible for non-smooth data; e.g., $\gamma < 0$. (The problems are an ill-conditioned weighting matrix for method of moments, no model solution for other methods.) These problems with the habit model are well documented in Aldrich and Gallant

(2011) for data that include the Great Depression and we find similar problems here because our data includes the Great Recession. For our data, there are no numerical problems when all $\tau_i \leq 0.10$ in the prior. Numerical problems commence when τ_i exceeds 0.30 for g and σ. Therefore, we impose $\tau_i \leq 0.30$ for g and σ. (Bayesian methods require exploration of the posterior over much of its support whereas frequentist methods require evaluation of the Laplace criterion (4) over a region local to the optimum. Thus, the region where the method of moments criterion must be numerically stable for the two methods can differ.)

Also, we impose the support conditions $-0.5 < g < 0.5$, $\sigma > 0$, $\sigma_w > 0$, $-1 < \rho < 1$, $-1 < \phi < 1$, $0.7 < \delta < 1.05$, and $1 < \gamma < 20$. Of these, the only one that is not innocuous is $\gamma > 1$, which binds and visibly truncates the posterior for $\tau_i > 0.50$.

The frequentist method of moments estimate using moment functions (32) through (43) shown below and with ϕ and δ set to 0.87 and 0.89, respectively, is $\hat{\theta} = (0.0214,\ 0.0110,\ 0.149,\ 0.0915,\ 0.87,\ 0.89,\ 1.055)$ with standard errors on the same order of magnitude as the standard deviations shown in Table 4 save for ϕ and δ, which have standard errors of zero. The condition number of the weighting matrix for parameter values local to the optimum of the Laplace criterion is not unreasonably large. The frequentist estimates that appear anomalous are ρ and γ.

In what follows, the τ_i take on the values 0.01, 0.10, 0.50, 1.00, 2.00 with the exceptions noted above for g, σ, ϕ, and δ. The likelihood is near enough to its plateau when $\tau_i = 2.00$ that one learns nothing more from the data for higher values of τ_i. As seen from Table 4, the model with $\tau_i = 0.5$, with exceptions as just noted, is preferred. The Campbell and Cochrane (1999) calibration can be rejected. The left truncation of the marginal posterior for γ is mild for $\tau_i = 0.5$

(Table 4 about here)

The moment conditions used for inference are

$$
\begin{aligned}
m_{1,t} &= c_t - c_{t-1} - g && (32)\\
m_{2,t} &= \sigma^2 - (c_t - c_{t-1} - g)^2 && (33)\\
m_{3,t} &= \sigma_w^2 - (d_t - d_{t-1} - g)^2 && (34)\\
m_{4,t} &= \rho - (c_t - c_{t-1} - g)(d_t - d_{t-1} - g)/(\sigma\sigma_w) && (35)\\
m_{5,t} &= 1.0 - M_t(P_{d,t} + D_t)/P_{d,t-1} && (36)\\
m_{6,t} &= 1.0 - M_t(P_{f,t} + I_t)/P_{f,t-1}) && (37)\\
m_{7,t} &= r_{d,t-1}m_{5,t} && (38)\\
m_{8,t} &= r_{f,t-1}m_{5,t} && (39)\\
m_{9,t} &= (\ell_{t-1} - \ell_{t-2})m_{5,t} && (40)\\
m_{10,t} &= r_{d,t-1}m_{6,t} && (41)\\
m_{11,t} &= r_{f,t-1}m_{6,t} && (42)\\
m_{12,t} &= (\ell_{t-1} - \ell_{t-2})m_{6,t} && (43)
\end{aligned}
$$

where ℓ_t is the log of income growth at time t. Let m_t denote the column vector with elements $m_{i,t}$ for $i = 1, \ldots, 12 = M$.

Moment functions (32) through (35) are textbook method of moments equations for estimating g, σ, σ_w, and ρ. As a practical matter ϕ and δ are not identified for the reasons discussed above. The identification of ϕ and δ comes from the prior (31). With ϕ and δ pinned down by the prior, the identification of λ and δ follows immediately from moment functions (36) through (43) which are the textbook moment equations for method of moments estimation of λ and δ for CRRA style utility.

In a sample of size n we compute

$$\bar{m}(x, \theta) = \frac{1}{n}\sum_{t=1}^{n} m_t \tag{44}$$

and the continuously updated, one lag HAC weighting matrix $W(x, \theta)$ with Parzen weights (Gallant, 1987, p. 446, 533). We assert that $\sqrt{n}[W(x, \theta)]^{-\frac{1}{2}}\bar{m}(x, \theta)$ is distributed as $N_M(0, I)$, which implies that the likelihood has the functional form

$$p(x \,|\, \theta) = (2\pi)^{-\frac{M}{2}} \exp\left[-\frac{n}{2}\bar{m}'(x, \theta)W^{-1}(x, \theta)\bar{m}(x, \theta)\right], \tag{45}$$

where x denotes the data observed over the period 1942–2013.

Specifically, x is comprised of $P_{d,t}$, $P_{f,t}$, D_t, and I_t from the CRSP (2013) series on value-weighted returns including and excluding dividends, and one year bond returns, deflated using the GDP deflator from BEA (2013). And comprised of real per-capita consumption C_t (non-durables plus services) and personal income ℓ_t, which are current dollar series from BEA (2013), deflated using the same GDP deflator. The years 1950–2013 ($n = 64$) were used for estimation with the years 1942–1949 used to provide lags for computing $\bar{m}(x, \theta)$ and $W(x, \theta)$.

MCMC was used for estimation. The MCMC chains were comprised of 100,000 draws well past the point where transients died off. The proposal was move-one-at-a-time random walk. Posterior model probabilities are computed using the Newton and Raftery (1994) $\hat{p}^4$ method for computing the marginal likelihood from an MCMC chain when assigning equal prior probability to each model. The software used is in the public domain and available together with a User's Guide at `http://www.aronaldg.org/webfiles/mle`. The code and data for the results here are one of the examples included in the distribution.

Results are shown in Table 4. As seen from the table, Campbell and Cochrane's (1999) calibration is assigned negligible posterior probability. The preferred model has $\tau_i = 0.5$. All parameter estimates drift as the prior is relaxed, the most interesting of which is the drift in γ from 2 to 3. The reason that the Bayes estimate of γ at $\tau_i = 2$ differs so markedly from the frequentist estimate $\hat{\theta}$ above is that the posterior for γ is right skewed so that the mean is to the right of the mode. The mode of the posterior is more analogous to the frequentist's estimator.

The main advantage of method of moments is that one does not have to solve a model in order to conduct inference. Nonetheless, it is of interest to know what the efficiency loss might be if one were willing to solve the model. Gallant and McCulloch (2009) proposed a Bayesian method that relies on the ability to simulate a model in order to synthesize a likelihood. This idea when coupled with a sieve likelihood provides an estimator that is nearly as efficient as the estimator that uses the exact likelihood, were it available. Aldrich and Gallant (2011) used this method with a sieve to fit the habit model to annual data from 1930–2008. The coefficients of variation in percent of their estimates for parameters

not pinned down by the prior are $\text{cv}(g, \sigma, \sigma_w, \rho, \gamma) = (4.7, 3.7, 4.4, 4.7, 3.9)$. Our prior with $\tau_i = 0.1$ conforms most closely to theirs. The coefficient of variation for our estimates (Table 4) are $\text{cv}(g, \sigma, \sigma_w, \rho, \gamma) = (3.2, 3.7, 4.9, 5.1, 5.2)$. Taking sample sizes into account, one would expect the Aldrich and Gallant coefficients of variation to be smaller by about a factor of $\sqrt{64/79} = 0.9$. Assuming that what is done here and what was done by Aldrich and Gallant represents best practice for both estimators, it seems that two methods are roughly equivalent with respect to efficiency.

5 Conclusion

We explored the consequences of an assertion that moment functions comprised of data and the parameters of a structural model follow a distribution. We concluded that the assertion implies a distribution on the constituents of the moment functions and permits Bayesian inference on model parameters. Specifically, if the moment functions have one of the properties of a pivotal, then the assertion of a distribution on moment functions coupled with a proper prior permits Bayesian inference. Without the semi-pivotal condition, the assertion of a distribution for moment functions either partially or completely specifies the prior. In this case Bayesian inference may or may not be practicable depending on how much of the distribution of the constituents remains indeterminate after imposition of a non-contradictory prior. An asset pricing example using data from the US economy was used to illustrate the ideas.

6 References

Aldrich, Eric M., and A. Ronald Gallant (2011), "Habit, Long-Run Risks, Prospect? A Statistical Inquiry," *Journal of Financial Econometrics* 9, 589–618.

BEA (2013), Bureau of Economic Analysis, "Table 7.1. Selected Per Capita Product and Income Series in Current and Chained Dollars," www.bea.gov.

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

Chernozhukov, Victor, and Han Hong (2003), "An MCMC Approach to Classical Estimation," *Journal of Econometrics* 115, 293–346.

CRSP (2013), Center for Research in Security Prices, Graduate School of Business, The University of Chicago, Used with permission. All rights reserved. www.crsp.uchicago.edu

Duan, Jason, and Carl F. Mela (2009), "The Role of Spatial Demand on Outlet Location and Pricing," *Journal of Marketing Research* 46, 260–278 .

Fisher, R. A. (1930), "Inverse Probability," *Proceedings of the Cambridge Philosophical Society 26,* 528–535.

Gallant, A. R. (1987), *Nonlinear Statistical Models,* New York: Wiley.

Gallant, A. Ronald, Raffaella Giacomini, and Giuseppe Ragusa (2014), "Generalized Method of Moments with Latent Variables," Working paper, Department of Economics, Penn State University, http://www.aronaldg.org/papers/liml.pdf

Gallant, A. Ronald, and Han Hong (2007), "A Statistical Inquiry into the Plausibility of Recursive Utility," *Journal of Financial Econometrics* 5, 523–590.

Gallant, A. R., and R. E. McCulloch (2009), "On the Determination of General Statistical Models with Application to Asset Pricing," *Journal of the American Statistical Association,* 104, 117–131.

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

Newton, M. A., and A. E. Raftery (1994), "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap," *Journal of the Royal Statistical Society (B)* 56, 3–48.

Table 1. Tossing two dice (X, Λ) when the probability of the difference $D = X - \Lambda$ is the primitive.

Preimage	d	$P(D = d)$	$P(D = d \mid \Lambda = 1)$	$P(D = d \mid \Lambda = 2)$
$C_{-5} = \{(1,6)\}$	-5	0	0	0
$C_{-4} = \{(1,5),(2,6)\}$	-4	0	0	0
$C_{-3} = \{(1,4),(2,5),(3,6)\}$	-3	0	0	0
$C_{-2} = \{(1,3),(2,4),(3,5),(4,6)\}$	-2	0	0	0
$C_{-1} = \{(1,2),(2,3),(3,4),(4,5),(5,6)\}$	-1	4/18	0	4/18
$C_0 = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$	0	10/18	10/14	10/18
$C_1 = \{(2,1),(3,2),(4,3),(5,4),(6,5)\}$	1	4/18	4/14	4/18
$C_2 = \{(3,1),(4,2),(5,3),(6,4))\}$	2	0	0	0
$C_3 = \{(4,1),(5,2),(6,3)\}$	3	0	0	0
$C_4 = \{(5,1),(6,2)\}$	4	0	0	0
$C_5 = \{(6,1)\}$	5	0	0	0

Source: Gallant, Giacomini, and Ragusa (2014)

Table 2. Tossing two dice (X, Λ) when the probability of $D = X - \Lambda$, $E = 2X - \Lambda$ is the primitive.

Preimage	d, e	$P(D, E = d, e)$	$P(D, E = d, e \mid \Lambda = 1)$	$P(D, E = d, e \mid \Lambda = 2)$
$C_{0,1} = \{(1,1)\}$	0, 1	1/6	1	0
$C_{-1,0} = \{(1,2)\}$	-1, 0	1/18	0	1/3
$C_{-2,-1} = \{(1,3)\}$	-2,-1	0	0	0
$C_{-3,-2} = \{(1,4)\}$	-3,-2	0	0	0
$C_{-4,-3} = \{(1,5)\}$	-4,-3	0	0	0
$C_{-5,-4} = \{(1,6)\}$	-5,-4	0	0	0
$C_{1,3} = \{(2,1)\}$	1, 3	0	0	0
$C_{0,2} = \{(2,2)\}$	0, 2	1/18	0	1/3
$C_{-1,1} = \{(2,3)\}$	-1, 1	1/18	0	0
$C_{-2,0} = \{(2,4)\}$	-2, 0	0	0	0
$C_{-3,-1} = \{(2,5)\}$	-3,-1	0	0	0
$C_{-4,-2} = \{(2,6)\}$	-4,-2	0	0	0
$C_{2,5} = \{(3,1)\}$	2, 5	0	0	0
$C_{1,4} = \{(3,2)\}$	1, 4	1/18	0	1/3
$C_{0,3} = \{(3,3)\}$	0, 3	1/18	0	0
$C_{-1,2} = \{(3,4)\}$	-1, 2	1/18	0	0
$C_{-1,1} = \{(3,5)\}$	-1, 1	0	0	0
$C_{-1,0} = \{(3,6)\}$	-1, 0	0	0	0
$C_{3,7} = \{(4,1)\}$	3, 7	0	0	0
$C_{2,6} = \{(4,2)\}$	2, 6	0	0	0
$C_{1,5} = \{(4,3)\}$	1, 5	1/18	0	0
$C_{0,4} = \{(4,4)\}$	0, 4	1/18	0	0
$C_{-1,3} = \{(4,5)\}$	-1, 3	1/18	0	0
$C_{-2,2} = \{(4,6)\}$	-2, 2	0	0	0
$C_{-4,9} = \{(5,1)\}$	-4, 9	0	0	0
$C_{-3,8} = \{(5,2)\}$	-3, 8	0	0	0
$C_{-2,7} = \{(5,3)\}$	-2, 7	0	0	0
$C_{-1,6} = \{(5,4)\}$	-1, 6	1/18	0	0
$C_{0,5} = \{(5,5)\}$	0, 5	1/18	0	0
$C_{-1,4} = \{(5,6)\}$	-1, 4	0	0	0
$C_{5,11} = \{(6,1)\}$	5,11	0	0	0
$C_{4,10} = \{(6,2)\}$	4,10	0	0	0
$C_{3,9} = \{(6,3)\}$	3, 9	0	0	0
$C_{2,8} = \{(6,4)\}$	2, 8	0	0	0
$C_{1,7} = \{(6,5)\}$	1, 7	1/18	0	0
$C_{0,6} = \{(6,6)\}$	0, 6	1/6	0	0

Table 3. Data Characteristics

Variable	Mean	Std. Dev.
log consumption growth	0.02183	0.01256
log dividend growth	0.02117	0.1479
ρ	0.2399	
log income growth	0.02175	0.01925
geometric stock return	0.04355	0.1736
geometric bond return	0.02044	0.02969

Data are real, annual, per capital consumption and income for the years 1950–2013 and real, annual stock and bond returns for the same years from BEA (2013) and CRSP (2013). ρ is the correlation between log consumption growth and log dividend growth.

Table 4. Parameter Estimates for the Habit Model

	Prior Scale									
	$\tau = 0.01$		$\tau = 0.1$		$\tau = 0.5$		$\tau = 1$		$\tau = 2$	
Parameter	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
g	0.01896	0.0001040	0.02086	0.0006600	0.02196	0.001098	0.02198	0.001109	0.02205	0.001175
σ	0.01486	7.509e-05	0.01204	0.0004407	0.01093	0.0006138	0.01096	0.0006441	0.01091	0.0006507
σ_w	0.1121	0.0005665	0.1152	0.005595	0.1260	0.01581	0.1296	0.01781	0.1301	0.01884
ρ	0.2000	0.0009874	0.2002	0.01023	0.2033	0.04839	0.2065	0.08464	0.2189	0.1249
ϕ	0.8676	0.004260	0.8187	0.03066	0.8337	0.03649	0.8329	0.03685	0.8339	0.03520
δ	0.8886	0.004502	0.8742	0.02799	0.8898	0.03248	0.8873	0.03449	0.8799	0.03442
γ	2.0001	0.0150	1.9979	0.1038	2.0536	0.4894	2.3679	0.8108	3.0291	1.2303
Model Prob.	0		0.0036		0.4023		0.3345		0.2597	

Data are real, annual, per capital consumption and income for the years 1950–2013 and real, annual stock and bond returns for the same years from BEA (2013) and CRSP (2013) that are used to form the moment functions (32) through (43) with years prior to 1950 used for lags. The likelihood given by (45) is an assertion that the average of these moment functions over the data is normally distributed with variance given by a one lag HAC weighting matrix with Parzen weights (Gallant, 1987, p. 446). The prior is given by (31) with scale τ as shown in the table. It is an independence prior that states that the marginal probability is 95% that a parameter is within $\tau \times 100\,\%$ of Campbell and Cochrane's (1999) calibrated values with the exceptions of ϕ and δ which are as shown for the first two panels and 0.1 for last three panels and g and σ which are as shown for the first two panels and 0.3 for the last three panels. The columns labeled mean and standard deviation are the mean and standard deviations of an MCMC chain (Gamerman and Lopes (2006), Chernozukov and Hong, 2003) of length 100,000 collected past the point where transients have dissipated. The proposal is move-one-at-a-time random walk. Posterior model probabilities are computed using the Newton and Raftery (1994) $\hat{p}^4$ method for computing the marginal likelihood from an MCMC chain when assigning equal prior probability to each model. The software and data for this example are at `http://www.aronaldg.org/webfiles/mle`.