Generalized Method of Moments with Latent Variables*

A. Ronald Gallant
Penn State University

Raffaella Giacomini
University College London

Giuseppe Ragusa
Luiss University

First draft: December 21, 2012
This draft: June 10, 2014

*Address correspondence to A. Ronald Gallant, P.O. Box 659, Chapel Hill NC 27514, USA, phone 919-428-1130; email aronldg@gmail.com.
We thank Gary Chamberlain, Whitney Newey, Frank Schorfheide, Neil Shephard and seminar participants at Harvard/MIT, Cambridge, Carlos III, Yale, Duke, Northern Illinois, UCSC, Penn State, Michigan State, Cleveland Fed, ULB for useful comments and suggestions. Raffaella Giacomini gratefully acknowledges financial support from the ESRC through the Centre for Microdata Methods and Practice grant RES-589-28-0001.

Abstract

The contribution of generalized method of moments (Hansen and Singleton, 1982) was to allow frequentist inference regarding the parameters of a nonlinear structural model without having to solve the model. Provided there were no latent variables. The contribution of this paper is the same. With latent variables.

Keywords and Phrases: Generalized Method of Moments, Latent Variables, Structural Models, Particle Filter

JEL Classification: C32, C36, E27

1 Introduction

We propose a generalized method of moments (GMM) estimator (Hansen and Singleton, 1982) for frequentist inference regarding the parameters of a nonlinear structural model that has dynamic latent variables. By latent variables we mean all endogenous and exogenous variables in the model that are not observed. Under the assumptions listed in Section 2, the estimator is consistent and asymptotically normally distributed.

Intuitively the problem we address is this: GMM works by using data that can be viewed as a draw (i.e., a sample) from the finite sample distribution implied by a model to approximate an unconditional expectation. We are missing the data on the latent variables. One possible remedy is to draw the latent variables from the conditional distribution of the latent variables given the observed variables. The particle filter is a standard method for drawing from a conditional distribution. To use it one needs both to be able to draw from the marginal distribution of the latent variables and to be able to evaluate the conditional density of the observed variables given the latent variables. In this paper, we assume that we can draw from the marginal and we show that, if the GMM criterion is asymptotically normally distributed, we can synthesize a conditional density that will generate a valid particle filter.

In the literature to which we contribute (cf. Flury and Shephard (2010), Fernandez-Villaverde and Rubio-Ramirez (2006)) the assumption that one can draw from the marginal distribution of the latent variables is standard. Our contribution is to be able to draw from the conditional distribution of the latent variables given the observed variables without knowledge of the conditional distribution of the observed variables given the latent variables.

The specifics of the estimator we propose are as follows: We assume enough knowledge of the transition density of the latent variables that we can draw a future latent variable given the past and the model's parameters. Under this assumption, we can define a Metropolis within Gibbs algorithm with Chernozhukov and Hong's (2003) Markov Chain Monte Carlo (MCMC) algorithm as the Metropolis step and Andrieu, Doucet, and Holenstein's (2010, Subsection 4.1) conditional particle filter algorithm as the Gibbs step. The result is an MCMC chain in the parameters. Parameter estimates and their standard errors are

computed from this MCMC chain.

The main attraction of the method we propose is that one does not have to solve the structural model. For partial equilibrium models this is crucial because, in general, there do not exist practicable alternatives.

We also expect that an important application for our results will be statistical inference regarding general equilibrium models in macroeconomic applications such as dynamic stochastic general equilibrium models (DSGE). For this class of models there are a variety of methods one might consider.

For analytically intractable models there are alternatives to what we propose but they all rely on being able to solve the model numerically. For instance, one can use perturbation methods to approximate the model, use the approximation to obtain an analytical expression for the likelihood, and then use some method of numerical integration such as particle filtering to eliminate the latent variables along the lines proposed by Fernandez-Villaverde and Rubio-Ramirez (2006). Or, one can solve the model only to the point of being able to simulate it and then use either simulated method of moments (SMM) (Duffy and Singleton, 1993) or efficient method of moments (EMM) (Gallant and Tauchen, 1996). These cites are the ones that we think readers will find most useful. They are not attributions. For attributions see the cited papers.

The main reason one might want to consider alternatives to these frequentist inference procedures is that one has misgivings about the quality of the numerical methods one has used to solve the structural model. For instance, perturbation methods such as linearization cause loss of information: they typically require dealing with singularity issues and with possible multiplicity of solutions (indeterminacy). Moreover, lower order expansions can lose important features of a model such as stochastic volatility (Bloom, 2009; Benigno, Benigno, Nisticó, 2012). A secondary reason is to avoid singularities in the measurement equation that can arise when using a likelihood based approach with particle filtering; see, e.g., Subsection 6.2.

The aforementioned frequentist strategies have Bayesian counterparts. A state-of-the-art Bayesian counterpart to Fernandez-Villaverde and Rubio-Ramirez (2006) is Flury and Shephard (2010). A Bayesian counterpart to EMM is Gallant and McCulloch (2009).

There is a Bayesian counterpart to GMM with latent variables, namely Gallant and Hong (2007). They exploit some differences between Bayesian and frequentist inference with the consequence that their approach does not directly accomodate frequentist inference. In this paper we extend their constructions, which can be traced back to fiducial inference (Fisher, 1930), to accomodate frequentist inference. The main issue is showing that the conditional density we construct can be used to generate draws from the conditional density of the latent variables given the observed variables. Once this is done, the remainder of the analysis can be accomplished by applying standard results.

2 Assumptions

Our assumptions are high level. The exception is Assumption 6 below, which is both important and easily checked. Our justification for avoiding low level detail is twofold. It would (i) be a routine, lengthy, tedious repetition of standard arguments of the sort found in Gallant and White (1987) and Chernozhukov and Hong (2003) and (ii) for interesting, nonlinear, structural models regularity conditions are impossible to check and are largely irrelevant in applications. While a negative implication, especially with respect to identification, is useful because it saves one from performing computations in vain, a positive implication is of little value. If one is using derivative-based hill climbing methods, then what matters is that optimizations converge to a limited number of isolated maxima (or minima) from many starts and the computed Hessian is well conditioned. If one is using MCMC based methods then what matters is that the chain mixes for relatively easily determined choices of tuning parameters.

ASSUMPTION 1 We require the existence of (but not complete knowledge of) a dynamic structural model that has parameters θ, a vector, that lie in a parameter space Θ. We denote the true but unknown value of the parameters by θ^o. We observe the history $X = (X_1, X_2, ..., X_T)$, a subset of the endogenous and exogenous variables in the model. We do not observe the variables in the model that remain: $\Lambda = (\Lambda_1, \Lambda_2, ..., \Lambda_T)$. These are the latent variables. Partial histories are denoted $X_{1:t} = (X_1, X_2, ..., X_t)$ and $\Lambda_{1:t} = (\Lambda_1, \Lambda_2, ..., \Lambda_t)$. ∎

Denote the probability measure determined by the structural model by P_θ^o and note

specifically that here θ is not necessarily equal to the true value θ^o but rather can be any value in the parameter space Θ. Define $p^o(\Lambda \,|\, X, \theta)$ to be the conditional density function of Λ given X that corresponds to P^o_θ. Let $p^o(\Lambda, \theta)$ and $p^o(\Lambda_{t+1} \,|\, \Lambda_t, \theta)$ denote the marginal and transition densities for Λ that correspond to P^o_θ, respectively. When we wish to emphasize that the observed data is meant and not some hypothetical value we write X^o and $X^o_{1:t}$.

ASSUMPTION 2 We assume that we can draw from the transition density of the dynamic latent variables $\Lambda_{t+1} \sim p^o(\Lambda_{t+1} \,|\, \Lambda_t, \theta)$. The transition density is assumed to be ergodic. ∎

Examples of latent variables that satisfy Assumption 2 and are routinely used in economics models are time-varying parameters, structural shocks, state-dependent parameters, and state-dependent factors.

Note that the functional form $p^o(\Lambda_{t+1} \,|\, \Lambda_t, \theta)$ implies that we can draw from the stationary density $p^o(\Lambda_t \,|\, \theta)$ by drawing from $p^o(\Lambda_{t+1} \,|\, \Lambda_t, \theta)$ with an arbitrary start Λ_0 and waiting for transients to die out.

The model can exhibit state dependence; e.g., Markov switching. If necessary to accommodate state dependence, one can modify the functional form of the transition density provided that ergodicity is retained because the only use made of the transition density is to propose a value of Λ_{t+1} for the purpose of extending $\Lambda_{1:t}$. Therefore, the transition density could, e.g., be of the form $p^o(\Lambda_{t+1} \,|\, \Lambda_{1:t},\, X_{1:t},\, \theta)$. However, in this case, one must provide some means to obtain an initial draw. One approach would be to use the method proposed by Gallant and Hong (2007, p. 536), which starts with a guess for Λ_0, draws from $p^o(\Lambda_{t+1} \,|\, \Lambda_{1:t},\, X^o_{1:t},\, \theta)$, recursively, and uses the draw at $t = T$ as the start for estimation.

When working with DSGE models one is used to thinking in terms of observables and states. That is not the dichotomy we have in mind here. Our division is into what is observed and what is not observed. Thus, what we term latent variables can include unobserved states, unobserved exogenous variables, and unobserved endogenous variables. The practical limit on what is permitted is determined by Assumption 2 (and the preceding paragraph).

Throughout we rely on conventional asymptotics, e.g., Hansen (1982), Gallant and White (1987), and Chernozhukov and Hong (2003), which rules out most unit root type behavior. This may require that a parameter lie in an interval, which is a condition that is trivially

easy to impose on one or more parameters at the Metropolis step of our proposed estimation method.

ASSUMPTION 3 We are given a set of conditional moment conditions of the form

$$\mathcal{E}[g(X_{t+1}, \Lambda_{t+1}, \theta) \,|\, \mathcal{I}_t] = 0,$$

where the information set is $\mathcal{I}_t = \{X_{-\infty}, ..., X_t \, \Lambda_{-\infty}, ..., \Lambda_t\}$. We assume that the unconditional moment conditions

$$0 = \mathcal{E}[g(X_{t+1}, \Lambda_{t+1}, \theta)] = \mathcal{E}\left[\int g(X_{t+1}, \Lambda_{t+1}, \theta)\, p^o(\Lambda \,|\, X, \theta^o)\, d\Lambda\right] \tag{1}$$

would identify θ if both X and Λ were observed. Similarly, we assume that the unconditional moment conditions

$$0 = \mathcal{E}[\bar{g}^*(X_{t+1}, \theta)] = \mathcal{E}\left[\int g(X_{t+1}, \Lambda_{t+1}, \theta)\, p^o(\Lambda \,|\, X, \theta)\, d\Lambda\right] \tag{2}$$

would identify θ if X were observed. ∎

Condition (1) is conventional, reasonably low level, and often not difficult to check. Its use is partial justification of (6). It would be the same as (2) were it not that the last θ appearing in (2) is evaluated at θ^o in (1). The reason for the extra condition (2) is that we do not observe Λ that is a draw from $p^o(\Lambda \,|\, X, \theta^o)$ as with conventional GMM but must use a draw from $p^o(\Lambda \,|\, X, \theta)$. This is entirely analogous to identification in maximum likelihood estimation with latent variables where one obtains a likelihood for X alone by integrating out Λ from the joint likelihood for X and Λ. It is the method of moments analogue of a standard assumption in the state space literature. Its use is partial justification of (7).

To perform a rough check on condition (2) one could integrate $g(X_{t+1}, \Lambda_{t+1}, \theta)$ with respect to the stationary distribution of $p^o(\Lambda_{t+1} \,|\, \theta^o)$ for putative θ^o to get an approximation to $\bar{g}^*(X_{t+1}, \theta)$. Alternatively, as mentioned above, one could rely on being able to get the MCMC chain that we propose to mix.

The method we propose, described in more detail below, consists of two steps: a Gibbs step that draws Λ given X, θ, and the previously drawn Λ; and, a Metropolis step that draws θ given X, Λ, and the previously drawn θ. We shall show that the θ draws are a sample from

the asymptotic distribution of the GMM estimator determined by (2) for large T. These draws are the means by which statistical inference is conducted. The moment conditions for the Gibbs and Metropolis steps can be different. For the Gibbs step the perfect moments would be those that spanned the scores of the the conditional density of X given Λ and θ, were they known. For the Metropolis step the perfect moments would be those that spanned the scores of the density for X given θ, were they known. These moment selection rules are not achievable in practice but they do provide guidance in applications. Another reason that one might want to split the moments into two groups is to reduce computation time. If, say, one can divide ten moment conditions into two groups of five each, then computation time would more than halve.

GMM estimation results depend on the skill one uses in constructing moment conditions. As just mentioned, by making sure that the moments used at the Metropolis step span the scores of the likelihood for observables (i.e., the density of X after eliminating Λ by integration), GMM results can be made the same as those for the maximum likelihood estimator (MLE), which are the best achievable. This is usually impossible without having an analytic expression for the likelihood, in which case there is no point to using GMM. However, there do seem to be some principles one can apply in selecting moments at the Metropolis step that we have discovered in our experimentation. One should try to identify as many parameters as possible from the observed data alone and try to make the latent variables depend as much as possible on quantities that can be computed from the observed data. If one is successful at this, then estimation results will be satisfactory, in our experience, but draws from the conditional distribution of the latent variables will not mimic the true (but unobserved) trajectory of the latent variables very well. This can be corrected, in our experience, by choosing the moments used in the Gibbs step so that observed variables depend on the latent variables as much as possible without regard for identification of parameters. I.e., the exact opposite of the goal for choosing moments for the Metropolis step. We illustrate these principles in the DSGE example of Subsection 6.2.

Oddly enough, if our DSGE example is not misleading, a poor choice of moments at the Gibbs step does not materially degrade the performance of the estimator for θ, as seen in Subsection 6.2. This is probably for two reasons: (i) The amount of noise in an unbiased

estimate of the likelihood in an MCMC chain affects the rejection rate of the chain and not much else; on this see Flury and Shephard (2010). (ii) The Gibbs algorithm has memory so that future draws will mimic the trajectory of the latent variables with high probability once a particle that mimics the trajectory has been drawn. This intuition probably explains why using a "likelihood" (cf., equation (10)) with $g^*(X_{t+1}, \theta)$ computed by averaging over a large number of particles rather than by evaluation at one Gibbs draw did not work well when we tried it. Averaging biases the estimate of the "likelihood" whereas our Gibbs strategy does not. Also, the Gibbs strategy is much cheaper to compute.

Some parameters of a model, particularly a DSGE model, may not be identified even if the correct likelihood involving only observables were known. This is a common problem in frequentist inference. When it occurs, the unidentified parameters must be calibrated or one must resort to methods for determining the boundaries of identified sets. Our DSGE example in Subsection 6.2 exhibits this problem and we deal with it by calibration.

Sample moment conditions corresponding to (1) are

$$g_T(X, \Lambda, \theta) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} g(X_t, \Lambda_t, \theta) \tag{3}$$

with weighting matrix

$$\Sigma(X, \Lambda, \theta) = \frac{1}{T} \sum_{t=1}^{T} \widetilde{g}(X_t, \Lambda_t, \theta)' \widetilde{g}(X_t, \Lambda_t, \theta) \tag{4}$$

$$\widetilde{g}(X_t, \Lambda_t, \theta) = g(X_t, \Lambda_t, \theta) - \frac{1}{\sqrt{T}} g_T(X, \Lambda, \theta) \tag{5}$$

If the moment conditions are serially correlated one will have to substitute a heteroskedastic autoregressive consistent (HAC) weighting matrix (Andrews, 1991) for that shown as (4). If a HAC matrix is used, the residuals used to compute it should be of the form shown as (5).

Define $g_T^*(X, \theta)$, $\Sigma^*(X, \theta)$, and $\widetilde{g}^*(X_t, \theta)$ the same as (3), (4), and (5), respectively, but with a draw from $p^o(\Lambda \mid X, \theta)$ replacing Λ throughout. It is important to note that the same draw is substituted for Λ in each of the equations (3), (4), and (5), not a different draw per equation.

ASSUMPTION 4 We assume that the sample moment conditions normalized by the

weighting matrix are asymptotically normal; i.e.,

$$Z = [\Sigma(X, \Lambda, \theta^o)]^{-1/2} g_T(X, \Lambda, \theta^o) \xrightarrow{d} N(0, I) \tag{6}$$

$$Z^* = [\Sigma^*(X, \theta^o)]^{-1/2} g_T^*(X, \theta^o) \xrightarrow{d} N(0, I), \tag{7}$$

and that the parameter space Θ is compact. ∎

Regularity conditions such that asymptotic normality obtains are in Hansen (1982), Gallant and White (1987), and elsewhere. Compactness of Θ is typically listed among these regularity conditions. We impose compactness here to provide a presumption that particle filter weights will be bounded and that the MCMC chain we propose will mix.

Define

$$p(X, \Lambda, \theta) = (2\pi)^{-M/2} \exp\left\{ -\frac{1}{2} g_T(X, \Lambda, \theta)' \left[\Sigma(X, \Lambda, \theta)\right]^{-1} g_T(X, \Lambda, \theta) \right\} \tag{8}$$

$$p^*(X, \Lambda, \theta) = p(X, \Lambda, \theta)\, p^o(\Lambda, \theta) \tag{9}$$

Note that (9) implies $p^*(X \mid \Lambda, \theta) = p(X, \Lambda, \theta)$. We show in Section 3 that (8) and (9) can be regarded as density functions that correspond to probability measures P_θ and P_θ^*, respectively. Also in Section 3, we show that the three measures assign the same probability to preimages of Z.

The GMM estimator that we are targeting has objective function

$$S_T^*(\theta) = g_T^*(X, \theta)' \left[\Sigma^*(X, \theta)\right]^{-1} g_T^*(X, \theta). \tag{10}$$

When we wish to emphasize that the GMM estimator to which we refer has objective function (10) we will use the notation GMM*.

ASSUMPTION 5 The Chernozhukov and Hong (2003) result holds for the target $S_T^*(\theta)$; that is, a sample $\left\{\theta^{(i)}\right\}_{i=1}^R$ from the density

$$p(\theta \mid X, \Lambda) \propto p(X, \Lambda, \theta) \tag{11}$$

with Λ a draw from $p^o(\Lambda \mid X, \theta)$ is a sample from the asymptotic distribution of the GMM* estimator for large T. ∎

This is the most high level of our assumptions. Inspection of the proofs in Chernozhukov and Hong (2003) suggests that Assumption 5 will be satisfied if the estimator that GMM* targets is strongly identified, consistent, and asymptotically normal. The method we propose is as follows:

1. Initialization. Choose a reasonable start $(\theta^{(0)},\ \Lambda^{(0)})$ and set $i = 1$.

2. Sample $\theta^{(i)}$ from $p(\theta \,|\, X, \Lambda^{(i-1)}) \propto p(X,\ \Lambda^{(i-1)},\ \theta)$ knowing $\theta^{(i-1)}$ using a Metropolis chain that starts at $\theta^{(i-1)}$. (Subsection 4.3).

3. Sample $\Lambda^{(i)}$ from $p^*(\Lambda \,|\, X, \theta^{(i)}) \propto p(X, \Lambda, \theta^{(i)}) p^o(\Lambda, \theta^{(i)})$ knowing $\Lambda^{(i-1)}$ using a particle filter that conditions on $\Lambda^{(i-1)}$ (Subsection 4.2).

4. Increment i and repeat from Step 2 until i exceeds some preassigned value R.

It is of interest in the frequentist context to be able to generate counterfactuals for Λ given some choice of X and θ. For this one needs the ordinary particle filter algorithm (Subsection 4.1) that generates draws from $p^*(\Lambda \,|\, X,\ \theta)$ without conditioning on a previous Λ draw. Actually, as we shall see in Section 5, it is only the ordinary particle filter algorithm that we have to derive because the particle filter that draws from $p^*(\Lambda \,|\, X, \theta)$ conditional on a previous draw of Λ obtains as a corollary.

Essentially what we propose is to use a Bayesian method, i.e., an MCMC chain, with an uninformative prior over compact Θ as a frequentist estimator. By the result of Chernozhukov and Hong (2003) the mode (or mean) of this chain is a consistent, asymptotically normal estimator with variance matrix estimated consistently by the variance matrix of the chain. This asymptotic distribution is that for target GMM*.

Also, we shall have to come to grips with the issue that the actual small sample distribution of Z is not the standard normal Φ on $\mathbb{R}^M$ but some other distribution Ψ_T, which issue we shall address in Sections 3 and 5. Excepting those two sections we shall ignore the distinction between Φ and Ψ_T because asymptotically it does not matter and we only use densities defined in terms of Φ and its density ϕ in all other sections of the paper.

We impose an additional requirement that is critical:

ASSUMPTION 6 Recall that $X_{1:t} = (X_1, \ldots, X_t)$, $\Lambda_{1:t} = (\Lambda_1, \ldots, \Lambda_t)$, and define

$$Z_t(X_{1:t}, \Lambda_{1:t}, \theta) = [\Sigma(X_{1:t}, \Lambda_{1:t}, \theta)]^{-1/2} g_t(X_{1:t}, \Lambda_{1:t}, \theta). \tag{12}$$

For each pair $(\Lambda_{1:t}, \theta)$ that the structural model permits, let $\mathcal{X}^{(\Lambda_{1:t},\theta)}$ be the set of permitted $X_{1:t}$. Let

$$C_z^{(\Lambda_{1:t},\theta)} = \{X_{1:t} \in \mathcal{X}^{(\Lambda_{1:t},\theta)} \,:\, Z_t(X_{1:t}, \Lambda_{1:t}, \theta) = z\}. \tag{13}$$

We assume that $C_z^{(\Lambda_{1:t},\theta)}$ is not empty for any $z \in \mathbb{R}^M$.

If $C_z^{(\Lambda_{1:t},\theta)}$ is not empty, then for each $(\Lambda_{1:t}, \theta)$ and z we may choose a point $X_{1:t}^* \in \mathcal{X}^{(\Lambda_{1:t},\theta)}$ for which

$$Z_t(X_{1:t}^*, \Lambda_{1:t}, \theta) = z.$$

Define

$$\Upsilon(z, \Lambda_{1:t}, \theta) = X_{1:t}^*. \tag{14}$$

Recall that $X_{1:t}^o$ denotes the observed $X_{1:t}$ and define $z^o = Z(X_{1:t}^o, \Lambda_{1:t}, \theta)$. For $z = z^o$ we shall choose the representer $X_{1:t}^*$ of $C_z^{(\Lambda_{1:t},\theta)}$ to be $X_{1:t}^o$ so that we always have $X_{1:t}^o = \Upsilon(Z_t(X_{1:t}^o, \Lambda_{1:t}, \theta), \Lambda_{1:t}, \theta)$.

If $C_z^{(\Lambda_{1:t},\theta)}$ is not empty, then the event

$$C_{\lambda_{1:t},z}^{\theta} = \{(X_{1:t}, \Lambda_{1:t}) \,:\, Z_t(X_{1:t}, \Lambda_{1:t}, \theta) = z,\ \Lambda_{1:t} = \lambda_{1:t}\}$$

can occur for every $z \in \mathbb{R}^M$. Then the union of all sets that can occur when $\Lambda_{1:t} = \lambda_{1:t}$ is known to have occurred is

$$O_{\lambda_{1:t}}^{\theta} = \cup_{z \in \mathbb{R}^M} C_{\lambda_{1:t},z}^{\theta}.$$

We assume $P_\theta^o(O_{\lambda_{1:t}}^{\theta}) = 1$. ∎

As yet we have not encountered a practical application that violates Assumption 6. Sufficient is that each element of g_t is unbounded and continuous with respect to at least one continuous element of X_t and that the residuals used to compute the weighting matrix are centered as in (5).

3 The Likelihood Induced by GMM

Gallant and Hong (2007) introduced a method for Bayesian inference for dynamic models with (possibly endogenous) unobserved variables building on ideas due to Fisher (1930) and used it to estimate the monthly and annual pricing kernels from a panel of equity and fixed income securities. In the course of this development they characterized the likelihood induced by GMM. We describe their ideas mostly through examples because our experience from comments on our work is that the ideas are easier to grasp from examples than by the general development in Gallant and Hong (2007). Subsection 3.5 concludes with summary of their ideas that relates directly to the examples couched in the notation of this paper.

This section makes the following points.

- A GMM criterion function induces a probability measure P_θ on a σ-algebra $\mathcal{C}$ containing sets C that have elements (X, Λ). Typically $\mathcal{C}$ is coarse in the sense that it does not contain all the Borel sets.

- Knowledge of the marginal distribution of Λ allows one to embed $\mathcal{C}$ within a σ-algebra $\mathcal{C}^*$ that contains the rectangles $R_B = \mathbb{R}^{\dim(X)} \times B$, where B is a Borel subset of $\mathbb{R}^{\dim(\Lambda)}$ and to define a probability measure P_θ^* on $\mathcal{C}^*$ that agrees with both P_θ on $\mathcal{C}$ and with the true data generating process P_θ^o on $\mathcal{C}$.

- The expectation $\mathcal{E}(f)$ of a $\mathcal{C}$-measurable function $f(X, \Lambda)$ has the same value whether computed under P_θ, P_θ^*, or P_θ^o.

In the examples θ is fixed so we can simplify by considering X and Λ only, leaving θ out of the discussion. The abstraction in Subsection 3.5 reintroduces θ.

3.1 A Probability Distribution Induced by GMM

Consider a situation where the probability distribution of a GMM criterion function D is known such as that shown in Table 1. In Table 1, D is the difference $D = X - \Lambda$ between tosses of two correlated, six-sided dice X and Λ. The expectation of D is zero.

(Table 1 about here)

In general, for a discrete probability measure defined on a measurable space $(\mathbb{R} \times \mathbb{R}, \mathcal{C})$, one conditions on knowing that a random variable Λ has the value λ by conditioning on the union of all sets in $\mathcal{C}$ that contain the point (x, λ) for some x. Denote this union by O_λ. O_λ is the union of all sets in $\mathcal{C}$ that can occur if $\Lambda = \lambda$ is known to have occurred. For the specific case shown in Table 1, the conditional probability density is

$$P(D = d \,|\, \Lambda = \lambda) = \frac{P(C_d \cap O_\lambda)}{P(O_\lambda)}, \tag{15}$$

where C_d is the preimage of d under D, as displayed in Table 1, and $\mathcal{C}$ is 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_λ is the rectangle $\mathbb{R} \times \{\lambda\}$, which in this example would reduce to $R_\lambda = \mathbb{D} \times \lambda$ with $\mathbb{D} = \{1, 2, 3, 4, 5, 6\}$. But in this example, $\mathcal{C}$ does not include the rectangles R_λ. If the σ-algebra over which probability is defined does not contain all the rectangles then O_λ need not take the form $\mathbb{R} \times \{\lambda\}$. Nonetheless, the principle expressed in (15) remains valid.

Because $P(C_d \cap O_\lambda) = \sum_{x=1}^{6} I_{C_d}(x, \lambda) P(D = d)$, an expression for $P(D = d \,|\, \Lambda = \lambda)$ is

$$P(D = d \,|\, \Lambda = \lambda) = \frac{\sum_{x=1}^{6} I_{C_d}(x, \lambda) P(D = d)}{\sum_{d=-5}^{5} \sum_{x=1}^{6} I_{C_d}(x, \lambda) P(D = d)}. \tag{16}$$

The denominator of (16) can be regarded as a "marginal" distribution

$$Q(\Lambda = \lambda) = P(O_\lambda) = \sum_{d=-5}^{5} \sum_{x=1}^{6} I_{C_d}(x, \lambda) P(D = d). \tag{17}$$

in the sense

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

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

$$\mathcal{E}(f \,|\, \Lambda = \lambda) = \sum_{x=1}^{6} f(x, \lambda) \sum_{d=-5}^{5} I_{C_d}(x, \lambda) P(D = d \,|\, \Lambda = \lambda) \tag{18}$$

can be used to compute conditional expectation because f can be regarded as a function of d and the right hand side of (18) equals

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

Equation (18) implies that we can view $P(D = d)$ as defining a conditional density function

$$P(X = x \,|\, \Lambda = \lambda) = \sum_{d=-5}^{5} I_{C_d}(x, \lambda) P(D = d \,|\, \Lambda = \lambda) \tag{19}$$

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

To get an expression that agrees with the expressions in Gallant and Hong (2007) note that we can write equation (19) as

$$P(X = x \,|\, \Lambda = \lambda) = \frac{P(D = x - \lambda)}{\sum_{x=1}^{6} P(D = x - \lambda)}. \tag{20}$$

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

The main idea in the development above is that if Λ is fixed at λ then d can be put into a one-to-one correspondence with x. Similar considerations define $P(D = d \,|\, X = x)$,

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

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

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 $P(X_1, X_2 = x_1, x_2 \,|\, \Lambda = \lambda)$. One way to circumvent the difficulty is as follows. Given λ, for each $s = 1, 2, \ldots, 6$ choose $(x_1, x_2) \in B_s$ to label B_s. The dominating measure puts mass one on these six representors and mass zero on all other pairs of integers.

The labeling of preimages by a representor X is one of the purposes of Assumption 6 although Assumption 6 is actually not satisfied this instance because we do not have $P(O_\lambda) = 1$. There are other ways to introduce a dominating measure so that $P(X_1, X_2 = x_1, x_2 \,|\, \Lambda = \lambda)$ can be regarded as a density. How one actually does it does not matter.

3.3 A Particle Filter for Distributions Induced by GMM

The particle filter is a recursive importance sampling scheme. Here we discuss the essential elements of the recursive step and an extension of P to a measure P^* on a larger σ-algebra using the example of the preceding subsection.

Assume that, in addition to the information in Table 1, the probability assigned to rectangles is known to be $P^*(R_\lambda) = \frac{1}{6}$. With this additional information we can augment the σ-algebra over which P is defined to include the rectangles R_λ. Let $\mathcal{C}^*$ denote the smallest σ-algebra that contains both $\{C_d\}_{d=-5}^{5}$ and $\{R_\lambda\}_{\lambda=1}^{6}$. 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 $\{(x, \lambda)\}$ are in $\mathcal{C}^*$ so that under P^* conditional probability has its conventional definition

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

Consider

$$\begin{aligned} &\sum_{\lambda=1}^{6} \frac{f(x,\lambda) P^*(X = x \,|\, \Lambda = \lambda)}{P^*(X = x)} P^*(\Lambda = \lambda) \\ &\qquad = \sum_{\lambda=1}^{6} \frac{f(x,\lambda) P^*(X = x, \Lambda = \lambda)}{P^*(\Lambda = \lambda) P^*(X = x)} P^*(\Lambda = \lambda) \qquad (21) \\ &\qquad = \sum_{\lambda=1}^{6} f(x,\lambda) P^*(\Lambda = \lambda \,|\, X = x) \end{aligned}$$

From equation (21) we observe that for $\mathcal{C}^*$-measurable f we can estimate

$$\mathcal{E}^*(f \,|\, X = x) = \sum_{\lambda=1}^{6} f(x,\lambda) P(\Lambda = \lambda \,|\, X = x)$$

by drawing a sample $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N$ from $P^*(\Lambda = \lambda)$ and estimating $\mathcal{E}^*(f \,|\, X = x)$ by

$$\tilde{\mathcal{E}}^*(f \,|\, X = x) = \frac{1}{N} \sum_{i=1}^{N} \frac{P^*(X = x \,|\, \Lambda = \tilde{\lambda}_i)}{P^*(X = x)} f(x, \tilde{\lambda}_i)$$

Because the constant function $f(x, \lambda) = 1$ is $\mathcal{C}^*$-measurable we have

$$1 = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \frac{P^*(X = x \,|\, \Lambda = \tilde{\lambda}_i)}{P^*(X = x)} \quad \text{a.s.}$$

from which it follows that if we do not know $P^*(X = x)$ then we can instead estimate $\mathcal{E}^*(f \,|\, X = x)$ by the weighted sum

$$\tilde{\mathcal{E}}^*(f \,|\, X = x) = \sum_{i=1}^{N} w_i f(x, \tilde{\lambda}_i), \tag{22}$$

where

$$w_i = \frac{\tilde{w}_i}{\sum_{i=1}^{N} \tilde{w}_i}$$

and

$$\tilde{w}_i = P^*(X = x \,|\, \Lambda = \tilde{\lambda}_i).$$

We can replace (22) by a formula with equal weights by resampling. That is, we view $\tilde{P}(\Lambda = \tilde{\lambda}_i) = w_i$ as defining a discrete probability distribution on the points $\{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_N\}$ and sample with replacement from $\tilde{P}$ to get $\hat{\lambda}_1, \ldots, \hat{\lambda}_N$ and estimate $\mathcal{E}^*(f \,|\, X = x)$ by

$$\hat{\mathcal{E}}^*(f \,|\, X = x) = \frac{1}{N} \sum_{i=1}^{N} f(x, \hat{\lambda}_i). \tag{23}$$

An implication of (23) is that we can view $\hat{\lambda}_1, \ldots, \hat{\lambda}_N$ as a sample from $P^*(\Lambda = \lambda \,|\, X = x)$. These are the essential ideas behind the particle filter. A particle filter is a recursive algorithm with equations (22) and (23) defining the recursive step.

A difficulty is that the information in Table 1 and the knowledge that $P^*(R_\lambda) = \frac{1}{6}$ is not enough to deduce $P^*(\{(x, \lambda\})$ because that knowledge and the knowledge that only sixteen $P^*(\{(x, \lambda\})$ can be non-zero (cf. Table 1), implies a singular system of nine equations in sixteen unknowns. There is one linear dependency that reduces the effective number of equations to eight.

$$\frac{4}{18} \;=\; \sum_{i=1}^{5} P^*(\{(i, i+1)\}) \tag{24}$$

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

For some purposes all solutions will be observationally equivalent and the choice of solution will not matter. Unfortunately the particle filter does depend upon choice of solution and there appears to be no logic that would cause one to favor one solution over another. This difficulty can be circumvented when $P(O_\lambda) = 1$ as seen in the next example, which is actually a discretized variant of Fisher (1930).

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

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 $Z(x,\lambda)$ are

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

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

$\lambda \in \mathbb{N}$. Therefore,

$$P(Z = z \,|\, \Lambda = \lambda) = \frac{P(C_z \cap O_\lambda)}{P(O_\lambda)} = P(C_z) = \frac{1-p}{1+p}\, p^{|z|}, \tag{25}$$

which does not depend on λ. Consequently,

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

using logic analogous to that leading to equation (20). The situation $P(O_\lambda) = 1$ seems to be what occurs most often in applications because Z is usually (at least asymptotically) pivotal.

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

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

The principal guiding this choice of solution to equations analagous to (24) 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 3.1, this choice was not available 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

$$P^*(Z = z) = \sum_{\lambda \in \mathbb{N}} P^*(X = z + \lambda, \Lambda = \lambda) = P(Z = z) \sum_{\lambda \in \mathbb{N}} P^*(R_\lambda) = P(Z = z). \tag{26}$$

The correct probability is assigned to rectangles because

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

Equations (26) and (25) imply that $P^*(Z = z \,|\, \Lambda = \lambda) = P(Z = z \,|\, \Lambda = \lambda)$.

3.5 The Abstraction

Let P^o_θ denote the denote the probability measure on the Borel subsets of $\mathbb{R}^{\dim(X)} \times \mathbb{R}^{\dim(\Lambda)}$ defined by the true data generating process. Hold θ fixed for the remainder of this subsection understanding that what follows is meant to hold for every θ in the parameter space Θ. Let

$\mathcal{C}$ be the smallest σ-algebra containing the preimages $C = \{(X, \Lambda) : Z(X, \Lambda, \theta) \in B\}$ where B ranges over the Borel subsets of $\mathbb{R}^M$ and Z is given by (6). Let $\mathcal{C}^*$ be the smallest σ-algebra that contains all sets in $\mathcal{C}$ plus all rectangles of the form $R_B = \mathbb{R}^{\dim(X)} \times B$, where B is a Borel subset of $\mathbb{R}^{\dim(\Lambda)}$. The σ-algebras $\mathcal{C}$ and $\mathcal{C}^*$ may depend on θ, which we let be understood to reduce notational clutter. Define $P_\theta(C) = P^o_\theta(C)$ for every $C \in \mathcal{C}$. Note that

$$(\mathbb{R}^{\dim(X)} \times \mathbb{R}^{\dim(\Lambda)}, \mathcal{C}, P^o_\theta) = (\mathbb{R}^{\dim(X)} \times \mathbb{R}^{\dim(\Lambda)}, \mathcal{C}, P_\theta)$$

so that expectations $\mathcal{E}(f)$are computed the same on each probability space for $\mathcal{C}$-measurable f. Assume that $Z(X, \Lambda, \theta)$ has distribution $\Psi_T(z)$ and density $\psi_T(z)$ under P^o_θ. $\psi_T(z)$ may depend on θ, which we let be understood; it will not if Z is pivotal with respect to P^o_θ. Let $p^o(\Lambda, \theta)$ be the marginal density for Λ implied by P^o_θ.

Recall that O_λ is the union of all sets in $\mathcal{C}$ that can occur if $\Lambda = \lambda$ is known to have occurred and note that Assumption 6 implies that the probability of O_λ is one under the probability space $(\mathbb{R}^{\dim(X)} \times \mathbb{R}^{\dim(\Lambda)}, \mathcal{C}, P_\theta)$. Motivated by the discussion in Subsection 3.4 we define

$$p^*(X \mid \Lambda, \theta) = p(X, \Lambda, \theta) = \psi_T[Z(X, \Lambda, \theta)]. \tag{27}$$

$$p^*(X, \Lambda, \theta) = p^*(X \mid \Lambda, \theta)\, p^o(\Lambda, \theta)$$

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

$$\int f(Z(x, \Lambda, \theta))\, p^*(x \mid \Lambda, \theta)\, dx = \int_{\mathbb{R}^{\dim(Z)}} f(z)\, \psi_T(z)\, dz, \tag{28}$$

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

$$\int I_B[Z(X, \Lambda, \theta)]\, p^*(x \mid \Lambda, \theta)\, dx = \int_B \psi_T(z)\, dz. \tag{29}$$

To each $C \in \mathcal{C}$ there is a Borel set B for which $C = \{(X, \Lambda) : Z(X, \Lambda, \theta) \in B\}$. Therefore, the fact that the right hand side of (29) does not depend on Λ implies that the probability measure P^*_θ that corresponds to the density $p^*(X, \Lambda, \theta)$ satisfies

$$P^*_\theta(C) = \int_B \psi_T(z)\, dz \tag{30}$$

for every $C \in \mathcal{C}$. For rectangles of the form $R_B = \mathbb{R}^{\dim(X)} \times B$ where B is a Borel subset of $\mathbb{R}^{\dim(\Lambda)}$ we already have that

$$P^*_\theta(R_B) = \int_B p^o(\lambda, \theta)\, d\lambda. \tag{31}$$

We conclude that P^*_θ and P^o_θ assign the same values to $C \in \mathcal{C}$ and to the rectangles R_B. For $C \in \mathcal{C}^*$ that cannot be computed using (30) and (31) 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

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

$$(\mathbb{R}^{\dim(X)} \times \mathbb{R}^{\dim(\Lambda)}, \mathcal{C}^*, P^o_\theta) = (\mathbb{R}^{\dim(X)} \times \mathbb{R}^{\dim(\Lambda)}, \mathcal{C}^*, P^*_\theta).$$

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.

For large T we have under Assumption 4 that $\Psi_T(z) \doteq \Phi(z)$ whence we can replace (27) by

$$p^*(X \,|\, \Lambda, \theta) = p(X, \Lambda, \theta) = \phi[Z(X, \Lambda, \theta)].$$

and drop the T from the marginal density for Λ to obtain expressions that agree with Section 2.

4 Algorithms

Three algorithms are required to implement our method:

- A particle filter (PF) algorithm.
  - Input: θ.
  - Output: Draws $\left\{\Lambda^{(i)}\right\}_{i=1}^N$ from $p^*(\Lambda \,|\, X, \theta)$.
- A Gibbs algorithm.
  - Input: The previous draw $\Lambda^{(i-1)}$ and a draw $\theta^{(i)}$ from $p(\theta \,|\, X, \Lambda^{(i-1)}) \propto p(X, \Lambda^{(i-1)}, \theta)$.
  - Output: A draw $\Lambda^{(i)}$ from $p^*(\Lambda \,|\, X, \theta^{(i)})$ that is conditional on $\Lambda^{(i-1)}$.
- A Metropolis algorithm.

– Input: The previous draw $\theta^{(i)}$ and a draw $\Lambda^{(i)}$ from $p^*(\Lambda \mid X, \theta^{(i)})$.
– Output: A draw $\theta^{(i+1)}$ from $p(\theta \mid X, \Lambda^{(i)}) \propto p(X, \Lambda^{(i)}, \theta)$ via a chain started at $\theta^{(i)}$.

In this section we present them in turn.

The particle filter algorithm produces draws from the conditional distribution of Λ given X and θ that is memoryless with respect to previous draws of Λ. The Gibbs algorithm is a corollary to the particle filter. It produces a draw from the conditional distribution of Λ given X and θ with memory of the previously drawn Λ, which improves computational efficiency. The Metropolis algorithm produces a draw from the conditional distribution of θ given Λ and X.

We previously introduced the notation $X_{1:t} = (X_1, ..., X_t)$ and $\Lambda_{1:t} = (\Lambda_1, ..., \Lambda_t)$ for partial histories. The joint density for partial histories is

$$p(X_{1:t}, \Lambda_{1:t}, \theta) = (2\pi)^{-M/2} \exp\left\{-\frac{1}{2} g_t(X_{1:t}, \Lambda_{1:t}, \theta)' \left[\Sigma(X_{1:t}, \Lambda_{1:t}, \theta)\right]^{-1} g_t(X_{1:t}, \Lambda_{1:t}, \theta)\right\}, \quad (32)$$

which corresponds to (8). The density $p^*(X_{1:t} \mid \Lambda_{1:t}, \theta)$ is equal to (32) and the density $p(\theta \mid \Lambda_{1:t}, X_{1:t})$ is proportional to (32). We do not need the proportionality factor for $p(\theta \mid \Lambda_{1:t}, X_{1:t})$ because we use a Metropolis algorithm to draw from it.

4.1 A Particle Filter

1. Initialization.

   - Input θ (and X)
   - Set T_0 to the minimum sample size required to compute $g_t(X_{1:t}, \Lambda_{1:t}, \theta)$.
   - For $i = 1, \ldots, N$ sample $(\Lambda_1^{(i)}, \Lambda_2^{(i)}, \ldots, \Lambda_{T_0}^{(i)})$ from $p^o(\Lambda_t | \Lambda_{t-1}, \theta)$.
   - Set t to $T_0 + 1$.
   - Set $\Lambda_{1:t-1}^{(i)} = (\Lambda_1^{(i)}, \Lambda_2^{(i)}, \ldots, \Lambda_{T_0}^{(i)})$

2. Importance sampling step.

   - For $i = 1, \ldots, N$ sample $\tilde{\Lambda}_t^{(i)}$ from $p^o(\Lambda_t | \Lambda_{t-1}^{(i)}, \theta)$ and set

$$\tilde{\Lambda}_{1:t}^{(i)} = (\Lambda_{0:t-1}^{(i)}, \tilde{\Lambda}_t^{(i)}).$$

- For $i = 1, \ldots, N$ compute weights $\tilde{w}_t^{(i)} = p^*(X_{1:t} \mid \tilde{\Lambda}_{1:t}^{(i)}, \theta)$.
- Scale the weights to sum to one.

3. Selection step.
   - For $i = 1, \ldots, N$ sample with replacement particles $\Lambda_{1:t}^{(i)}$ from the set $\{\tilde{\Lambda}_{1:t}^{(i)}\}$ according to the weights.

4. Repeat
   - If $t < T$, increment t and go to Importance sampling step;
   - else output $\left\{\Lambda_{1:T}^{(i)}\right\}_{i=1}^{N}$.

4.2 A Gibbs Algorithm

1. Initialization.
   - Input $\Lambda_{1:T}^{(1)}$, θ (and X)
   - Set T_0 to the minimum sample size required to compute $g_t(X_{1:t}, \Lambda_{1:t}, \theta)$.
   - For $i = \mathbf{2}, \ldots, N$ sample $(\Lambda_1^{(i)}, \Lambda_2^{(i)}, \ldots, \Lambda_{T_0}^{(i)})$ from $p^o(\Lambda_t | \Lambda_{t-1}, \theta)$.
   - Set t to $T_0 + 1$.
   - Set $\Lambda_{1:t-1}^{(i)} = (\Lambda_1^{(i)}, \Lambda_2^{(i)}, \ldots, \Lambda_{T_0}^{(i)})$

2. Importance sampling step.
   - For $i = \mathbf{2}, \ldots, N$ sample $\tilde{\Lambda}_t^{(i)}$ from $p_o(\Lambda_t | \Lambda_{t-1}^{(i)}, \theta)$ and set

$$\tilde{\Lambda}_{1:t}^{(i)} = (\Lambda_{0:t-1}^{(i)}, \tilde{\Lambda}_t^{(i)}).$$

   - For $i = \mathbf{1}, \ldots, N$ compute weights $\tilde{w}_t^{(i)} = p(X_{1:t} \mid \tilde{\Lambda}_{1:t}^{(i)}, \theta)$.
   - Scale the weights to sum to one.

3. Selection step.

- For $i = \mathbf{2}, \ldots, N$ sample with replacement particles $\Lambda_{1:t}^{(i)}$ from the set $\{\tilde{\Lambda}_{1:t}^{(i)}\}_{i=\mathbf{1}}^{N}$ according to the weights.

4. Repeat

- If $t < T$, increment t and go to Importance sampling step;
- else output the particle $\Lambda_{1:T}^{(N)}$.

4.3 A Metropolis Algorithm

To implement a Metropolis algorithm we require a proposal density for θ. A proposal density is a transition density of the form $T(\theta_{old}, \theta_{new})$ such as a move-one-at-a-time random walk. In the examples of Section 6, we used the move-one-at-a-time random walk that uniformly selects an index k and then moves the element $\theta_{k,old}$ of θ_{old} to $\theta_{k,new}$ according to a normal with mean $\theta_{k,old}$ and variance σ_k, where σ_k is chosen by trial and error to achieve a rejection rate of about 50% in the Accept-Reject step of the algorithm that follows. For K below we set $K = 50$ in our examples.

- Input: Λ, θ_{old} (and X)
- Propose: Draw θ_{prop} from $T(\theta_{old}, \theta)$
- Accept-Reject: Put $\theta^{(i)}$ to θ_{prop} with probability

$$\alpha = \min\left[1, \frac{p(X, \Lambda, \theta_{prop})T(\theta_{prop}, \theta_{old})}{p(X, \Lambda, \theta_{old})T(\theta_{old}, \theta_{prop})}\right]$$

  else put $\theta^{(i)}$ to θ_{old}.

- Repeat: If $i < K$ put $\theta_{old} = \theta^{(i)}$ and go to Propose; else output $\theta^{(K)}$.

5 Theory

5.1 Particle Filter Theory

THEOREM 1 Under Assumptions 1 through 6, the particle filter algorithm defined in Subsection 4.1 generates draws from $p^*(\Lambda \mid X, \theta)$.

Proof Refer to Assumption 6 for the notation which follows.

From the point of view of particle filter theory we have a transition density $p^o(\Lambda_t \,|\, \Lambda_{t-1}, \theta)$ and a measurement density

$$p^{\#}(z_t \,|\, \Lambda_{1:t}, \theta) = n\,\{[Z_t[\Upsilon(z_t, \Lambda_{1:t}, \theta), \Lambda_{1:t}, \theta] \,|\, 0, I\} \tag{33}$$

Note particularly that with θ and $\Lambda_{1:t}$ held fixed, the measurement density depends only on $z_t \subset \mathbb{R}^M$, $\Lambda_{1:t}$, and θ; it does not depend on $X_{1:t}$. The particle filter produces draws $\Lambda_{1:T}^{(i)}$ from the density $p^{\#}(\Lambda_{1:T} \,|\, z_{1:T}, \theta) \propto p^{\#}(z_t \,|\, \Lambda_{1:t}, \theta) p^o(\Lambda_{1:t}, \theta)$. However, Assumption 6 puts $z_{1:T}$ into a one-to-one correspondence with the observed $X_{1:T}^o$. Therefore,

$$p^*(\Lambda_{1:T} \,|\, X_{1:T}^o, \theta) = p^{\#}(\Lambda_{1:T} \,|\, z_{1:T}, \theta).$$

What we want are draws from the actual conditional density of $\Lambda = \Lambda_{1:T}$ given $X_{1:T}^o$ that we denote by $f_T(\Lambda \,|\, z_{1:T}, \theta)$. Let $\Psi_T(\cdot)$ denote the actual distribution of $Z_T(X_{1:T}^o, \Lambda, \theta)$ and $\psi_T(\cdot)$ its density function. We have assumed that $\Psi_T(\cdot)$ converges in distribution to the standard normal distribution $\Phi(\cdot)$, with density $\phi(\cdot)$, for large T. Let

$$u_T^{(i)} = \phi(z_T^{(i)})\, p^o(\Lambda \,|\, \theta) \tag{34}$$

$$U_T = \int \phi(Z_T(X_{1:T}^o, \Lambda, \theta))\, p^o(\Lambda \,|\, \theta)\, d\Lambda \tag{35}$$

$$v_T^{(i)} = \psi_T(z_T^{(i)})\, p^o(\Lambda \,|\, \theta) \tag{36}$$

$$V_T = \int \psi_T(Z_T(X_{1:T}^o, \Lambda, \theta))\, p^o(\Lambda \,|\, \theta)\, d\Lambda \tag{37}$$

where

$$p^o(\Lambda|\theta) = p^o(\Lambda_1^{(i)} \,|\, \theta) \prod_{s=2}^{T} p^o(\Lambda_s^{(i)} \,|\, \Lambda_{s-1}^{(i)}, \theta).$$

Using (34) through (37) to construct importance sampling weights, we have

$$\frac{1}{N} \sum_{i=1}^{N} \frac{v_T^{(i)}}{u_T^{(i)}} \frac{U_T}{V_T}\, g_T(X^o e_{:1:T}, \Lambda_{1:T}^{(i)}, \theta) = \frac{U_T}{V_T} \frac{1}{N} \sum_{i=1}^{N} \frac{\psi_T(z_T^{(i)})}{\phi(z_T^{(i)})}\, g_T(X_{1:T}^o, \Lambda_{1:T}^{(i)}, \theta) \tag{38}$$

is an approximation to

$$\int g_T(X_{1:T}^o, \Lambda, \theta)\, f_T(\Lambda \,|\, z_{1:T}, \theta)\, d\Lambda \tag{39}$$

The approximation error decreases as $N \to \infty$.

We shall first show that

$$\frac{U_T}{V_T}\frac{1}{N}\sum_{i=1}^{N} g_T(X^o_{1:T}, \Lambda^{(i)}_{1:T}, \theta) \tag{40}$$

also approximates (39) for large N and T.

Choose the cube $(a_0, b_0]$ large enough that

$$\frac{U_T}{V_T}\int I\{Z_T(X^o_{1:T}, \Lambda, \theta) \in (a_0, b_0]\}\, g_T(X^o_{1:T}, \Lambda, \theta)\, f_T(\Lambda \,|\, z_{1:T}, \theta)\, d\Lambda \tag{41}$$

approximates (39) to within $\epsilon/4$. Let $\eta = \min\{\phi(z) \,|\, z \in (a_0, b_0]\}$. The assumption of convergence in distribution implies that the convergence of $\Psi_T((a,b])$ to $\Phi((a,b])$ is uniform over all cubes of the form $(a,b]$ (Billingsly and Topsoe, 1967). Choose T large enough that $|\Psi_T((a,b]) - \Phi((a,b])| < \epsilon\eta/4$. Choose N large enough that

$$\frac{U_T}{V_T}\frac{1}{N}\sum_{i=1}^{N} I\{Z_T(X^o_{1:T}, \Lambda, \theta) \in (a_0, b_0]\}\, \frac{\psi_T(z_T^{(i)})}{\phi(z_T^{(i)})}\, g_T(X^o_{1:T}, \Lambda^{(i)}_{1:T}, \theta) \tag{42}$$

approximates (41) to within $\epsilon/4$. Choose cubes of the form $(a_i, b_i]$ of equal edge length h small enough that $\frac{\Psi_T((a_i,b_i])/h^M}{\Phi((a_i,b_i])/h^M}$ approximates $\frac{\psi_T(z_T^{(i)})}{\phi(z_T^{(i)})}$ to within $\epsilon/4$. We have shown that (40) approximates (39) to within ϵ.

We shall now show that $\frac{U_T}{V_T}$ tends to one.

Choose J disjoint rectangles $I_j = (c_j, d_j]$, where elements of c_j may be $-\infty$ and elements of d_j may be ∞, whose union is $\mathbb{R}^M$ and choose points $e_j \in I_j$ such that

$$\left|\sum_{j=1}^{J} \psi_T(e_j) I_{I_j}(z) - \psi_T(e_j)\right| < \epsilon$$

$$\left|\sum_{j=1}^{J} \phi(e_j) I_{I_j}(z) - \phi(e_j)\right| < \epsilon.$$

Note that $1 = \sum_{j=1}^{J} \int I_{I_j}(Z_T(X^o_{1:T}, \Lambda, \theta))\, p^o(\Lambda|\theta)\, d\Lambda$. Then for any T,

$$\begin{aligned}
&\frac{\sum_{j=1}^{J} \psi_T(e_j) \int I_{I_j}(Z_T(X^o_{1:T}, \Lambda, \theta))\, p^o(\Lambda|\theta)\, d\Lambda - \epsilon}{\sum_{j=1}^{J} \phi(e_j) \int I_{I_j}(Z_T(X^o_{1:T}, \Lambda, \theta))\, p^o(\Lambda|\theta)\, d\Lambda + \epsilon} \\
&\qquad < \frac{U_T}{V_T} \\
&\qquad < \frac{\sum_{j=1}^{J} \psi_T(e_j) \int I_{I_j}(Z_T(X^o_{1:T}, \Lambda, \theta))\, p^o(\Lambda|\theta)\, d\Lambda + \epsilon}{\sum_{j=1}^{J} \phi(e_j) \int I_{I_j}(Z_T(X^o_{1:T}, \Lambda, \theta))\, p^o(\Lambda|\theta)\, d\Lambda - \epsilon}
\end{aligned}$$

Choose cubes of the form $(a_j, b_j]$ of equal edge length h small enough that $\Psi_T((a_j, b_j])/h^M$ approximates $\psi_T(e_j)$ to within ϵ and $\Phi((a_j, b_j])/h^M$ approximates $\phi(e_j)$ to within ϵ, whence

$$\begin{aligned}
&\frac{\sum_{j=1}^J \Psi_T((a_j,b_j]) \int I_{I_j}(Z_T(X^o_{1:T},\Lambda,\theta))\, p^o(\Lambda|\theta)\, d\Lambda - 2\epsilon h^M}{\sum_{j=1}^J \Phi((a_j,b_j]) \int I_{I_j}(Z_T(X^o_{1:T},\Lambda,\theta))\, p^o(\Lambda|\theta)\, d\Lambda + 2\epsilon h^M} \\
&\qquad < \quad \frac{U_T}{V_T} \\
&\qquad < \quad \frac{\sum_{j=1}^J \Psi_T((a_j,b_j]) \int I_{I_j}(Z_T(X^o_{1:T},\Lambda,\theta))\, p^o(\Lambda|\theta)\, d\Lambda + 2\epsilon h^M}{\sum_{j=1}^J \Phi((a_j,b_j]) \int I_{I_j}(Z_T(X^o_{1:T},\Lambda,\theta))\, p^o(\Lambda|\theta)\, d\Lambda - 2\epsilon h^M}
\end{aligned}$$

Choose T large enough that $|\Psi_T((a,b]) - \Phi((a,b])| < \epsilon$, whence

$$\begin{aligned}
&\frac{\sum_{j=1}^J \Phi((a_j,b_j]) \int I_{I_j}(Z_T(X^o_{1:T},\Lambda,\theta))\, p^o(\Lambda|\theta)\, d\Lambda - \epsilon - 2\epsilon h^M}{\sum_{j=1}^J \Phi((a_j,b_j]) \int I_{I_j}(Z_T(X^o_{1:T},\Lambda,\theta))\, p^o(\Lambda|\theta)\, d\Lambda + \epsilon + 2\epsilon h^M} \\
&\qquad < \quad \frac{U_T}{V_T} \\
&\qquad < \quad \frac{\sum_{j=1}^J \Phi((a_j,b_j]) \int I_{I_j}(Z_T(X^o_{1:T},\Lambda,\theta))\, p^o(\Lambda|\theta)\, d\Lambda + \epsilon + 2\epsilon h^M}{\sum_{j=1}^J \Phi((a_j,b_j]) \int I_{I_j}(Z_T(X^o_{1:T},\Lambda,\theta))\, p^o(\Lambda|\theta)\, d\Lambda - \epsilon - 2\epsilon h^M}
\end{aligned}$$

which proves that $\frac{U_T}{V_T}$ tends to one.

Regularity conditions sufficient to imply that particles are draws from the density $p^{\#}(\Lambda_{1:T} \,|\, z_{1:T}, \theta)$ are in Andrieu, Doucet, and Holenstein (2010). They are mild, requiring that the weights at the importance sampling step be bounded and that multinomial resampling be used, which is the scheme used at the selection step.

The regularity conditions used to prove consistency and asymptotic normality of GMM estimators typically include a compact parameter space, domination conditions on the moment conditions, and bounds on the eigenvalues of the weighting matrix so that bounded weights are typically a side effect of these conditions. □

5.1.1 Comments on Particle Filter Theory

The performance of the particle filter depends upon the variance of the weights. As remarked earlier, on can use penalty functions to help in this regard. However, even with a penalty function, for small t there are few degrees of freedom for computing the weighting matrix and the variance of the weights is a problem. One might try to control this by setting T_0 larger than strictly necessary at the initialization step of the particle filter in Section 4.1

but doing this has a deleterious effect on the performance of the particle filter because the information from X is not being used until t exceeds T_0.

A better approach is regularization of the weighting matrix. If the condition number of the weighting matrix (ratio of smallest singular value value to the largest) falls below a preset value η (e.g. $\eta = 10^{-8}$) an amount δ is added to the diagonal elements of the weighting matrix just sufficient to bring the condition number to η prior to inversion of the weighting matrix.

5.2 Gibbs Theory

The proof above that we can draw a sample from $f_T(\Lambda \mid z_{1:T}, \theta)$ with negligible error for large T implies that the algorithm given in Subsection 4.1 of Andrieu, Doucet, and Holenstein (2010) is valid. This, in turn, implies that the algorithm proposed in Subsection 4.2 generates a valid Gibbs draw under the setup defined by Assumptions 1 through 6.

5.2.1 Comments on Gibbs Theory

Using only one particle to evaluate the conditional expectation is essential. One is relying on egodicity to correctly evaluate g^* and relying on the particle filter to provide an unbiased estimate of GMM*. If one averaged over several particles to compute g^* one would destroy the unbiasedness in the computation GMM*. On this see Andrieu, Doucet, and Holenstein (2010).

As to the number of particles one should use in the conditional particle filter, we found that $N = 1000$ gave about the same results as $N = 5000$ and larger. Andrieu, Doucet, and Holenstein (2010) report similar experience for their examples and suggest that the length of the MCMC chain R be increased rather than N because runtimes increase less with R than with N for most of their examples. Because our runtimes increase at the rate $RM[(T!)N + TK]$, the suggestion that N be kept small at the cost of increasing R carries considerable force.

5.3 Metropolis Theory

A compact parameter space, an identified model, and a move-one-at-a-time proposal are enough to ensure that Metropolis part of the Metropolis within Gibbs algorithm will mix (Gamerman and Lopes, 2006).

6 Examples

We illustrate our proposal with two examples: a stochastic volatility model with comparison to the Flury and Shephard (2010) estimator, and a DSGE model with comparision to the maximum likelihood estimator.

6.1 A Stochastic Volatility Model

Our first example is a stochastic volatility (SV) model:

$$\begin{aligned} X_t &= \rho X_{t-1} + \exp(\Lambda_t)\, u_t \\ \Lambda_t &= \phi \Lambda_{t-1} + \sigma e_t \\ e_t &\sim N(0,1) \\ u_t &\sim N(0,1) \end{aligned}$$

The true values of the parameters are

$$\theta_0 = (\rho_0, \phi_0, \sigma_0) = (0.9, 0.9, 0.5)$$

for the purpose of plotting the particle filter and

$$\theta_0 = (\rho_0, \phi_0, \sigma_0) = (0.25, 0.8, 0.1)$$

for illustrating estimation results. The reason for the difference is that the former generates plots that are easy to assess visually whereas the latter are more representative of, say, fits to daily S&P 500 closing prices.

The moment conditions used with this model are:

$$g_1 = (X_t - \rho X_{t-1})^2 - [\exp(\Lambda_t)]^2 \tag{43}$$

$$g_2 = |X_t - \rho X_{t-1}||X_{t-1} - \rho X_{t-2}| - \left(\frac{2}{\pi}\right)^2 \exp(\Lambda_t)\exp(\Lambda_{t-1}) \quad (44)$$

$$\vdots$$

$$g_{L+1} = |X_t - \rho X_{t-1}||X_{t-L} - \rho X_{t-L-1}| - \left(\frac{2}{\pi}\right)^2 \exp(\Lambda_t)\exp(\Lambda_{t-L}) \quad (45)$$

$$g_{L+2} = X_{t-1}(X_t - \rho X_{t-1}) \quad (46)$$

$$g_{L+3} = \Lambda_{t-1}(\Lambda_t - \phi\Lambda_{t-1}) \quad (47)$$

$$g_{L+4} = (\Lambda_t - \phi\Lambda_{t-1})^2 - \sigma^2 \quad (48)$$

Moment (46) identifies ρ independently of Λ_t; moments (43) through (46) overidentify Λ_t given ρ. Moment (47) identifies ϕ given Λ_t and moment (48) identifies σ given Λ_t and ϕ.

What may not be obvious here is how an equation such as (43) identifies Λ_t . One can see this at the point at which one computes weights in the importance sampling step of the PF algorithm (Subsection 4.1). The weight w_t depends on Λ_t while the weight w_{t-1} does not. Therefore the incremental information regarding Λ_t provided by (43) does get used at time t to determine Λ_t. For the Metropolis within Gibbs algorithm itself, the incremental information does get used at the Gibbs step but does not get used at the Metropolis step because the Metropolis step uses sums over all the data rather than partial sums.

Sometimes one uses a penalty function in connection with MCMC. We shall investigate the effect of multiplying (8) by the Jacobian term $[\det \Sigma(X, \Lambda, \theta)]^{-M/2}$. The idea is that (8) bases inference on the density for Z whereas (8) multiplied by the Jacobian term bases inference on the density for g_T. It is interesting to see what effect this might have.

Estimates of θ for the SV model are shown in Table 2 for three methods: Metropolis within Gibbs GMM with a Jacobian term, without a Jacobian term, and using the Flury and Shephard (2010) estimator. The Flury and Shephard estimator can be regarded as state-of-the-art. The MCMC chain generated using the method are draws from the exact posterior with a flat prior.

Applying the particle filter at the true value of θ and $N = 5000$, we obtain the estimate of Λ shown as a time series plot in Figure 1 and as a scatter plot in Figure 2 for the case when a Jacobian term is included and as Figures 3 and 4 when it is not. The plots for the Flury and Shephard estimator are Figures 5 and 6. In the particle filter vernacular, the

Metropolis within Gibbs GMM estimator is computed from a smooth whereas the Flury and Shephard estimator is computed from a filter; accordingly, the plots shown for the Metropolis within Gibbs GMM estimator are smooths whereas the plots shown of the Flury-Shephard estimator are filters.

(Table 2 about here)

(Figure 1 about here)

(Figure 2 about here)

(Figure 3 about here)

(Figure 4 about here)

(Figure 5 about here)

(Figure 6 about here)

6.2 A Dynamic Stochastic General Equilibrium Model

The second example is taken from Del Negro and Schorfheide (2008). We need to have a model with an exact analytical solution to generate accurate data with which to test our proposed methods. The working paper version of the article has some simplified versions of the full model in the article that have an analytic expression for the solution. The example is one of the simplified versions.

The full model is a medium-scale New Keynesian model with price and wage rigidities, capital accumulation, investment adjustment costs, variable capital utilization, and habit formation. The simplified model discussed here is obtained by removing capital, fixed costs, habit formation, the central bank, and making wages and prices flexible. With these choices, the model has three shocks: the log difference of total factor productivity z_t, a preference shock that affects intertemporal substitution between consumption and leisure ϕ_t, and the price elasticity of intermediate goods λ_t, called a mark-up shock in the article. In the full model the endogenous variables are output, consumption, investment, capital, and the real

wage, which are detrended by $\exp(z_t)$ and expressed as log deviations from the steady-state solution of the model, and inflation. Of these, the ones of interest in the simplified model are the log deviations of wages and output, w_t and y_t, respectively, and inflation π_t. The time increment is one quarter.

The exogenous shocks are

$$\begin{aligned} z_t &= \rho_z z_{t-1} + \sigma_z \epsilon_{z,t} \\ \phi_t &= \rho_\phi \phi_{t-1} + \sigma_\phi \epsilon_{\phi,t} \\ \lambda_t &= \rho_\lambda \lambda_{t-1} + \sigma_\lambda \epsilon_{\lambda,t}, \end{aligned} \tag{49}$$

where $\epsilon_{z,t}$, $\epsilon_{\phi,t}$, and $\epsilon_{\lambda,t}$ are independent standard normal random variables.

The first order conditions are

$$\begin{aligned} 0 &= y_t + \frac{1}{\beta}\pi_t - \mathcal{E}_t(y_{t+1} + \pi_{t+1} + z_{t+1}) \\ 0 &= w_t + \lambda_t \\ 0 &= w_t - (1+\nu) y_t - \phi_t \end{aligned} \tag{50}$$

where ν is the inverse Frisch labor supply elasticity and β is the subjective discount rate.

The solution for the endogenous variables is

$$\begin{aligned} w_t &= -\lambda_t \\ y_t &= -\frac{1}{1+\nu}\lambda_t - \frac{1}{1+\nu}\phi_t \\ \pi_t &= \beta \frac{1-\rho_\lambda}{(1+\nu)(1-\beta\rho_\lambda)}\lambda_t + \beta \frac{1-\rho_\phi}{(1+\nu)(1-\beta\rho_\phi)}\phi_t + \beta \frac{\rho_z}{(1-\beta\rho_z)} z_t \end{aligned} \tag{51}$$

The true values of the parameters are

$$\theta = (\rho_z, \rho_\phi, \rho_\lambda, \sigma_z, \sigma_\phi, \sigma_\lambda, \nu, \beta) = (0.15, 0.68, 0.56, 0.71, 2.93, 0.11, 0.96, 0.996)$$

which are the parameter estimates for model $\mathcal{P}_S$ of Del Negro and Schorfheide (2008) as supplied by Frank Schorfheide in an email communication.

We take w_t, y_t, and π_t as measured and z_t and ϕ_t as latent so that in our notation

$$X_t = (w_t, y_t, \pi_t)$$

$$\Lambda_t = (z_t, \phi_t).$$

This model is simple enough that an analytical expression for the likelihood is immediately available by substituting equations (49) into equations (51). By inspection one can anticipate identifications issues: a small change in σ_ϕ can be compensated by small changes to ν, β, and σ_z. This in turn, causes the MCMC chain for estimating the model by maximum likelihood (Chernozhukov and Hong, 2003) to fail to mix. If one is going to estimate this model by frequentist methods, one must, as a practical matter, calibrate three of the four parameters σ_z, σ_ϕ, ν, and β. Our choice is to calibrate σ_z, σ_ϕ, and ν, leaving β as the free parameter. The situation here is rather stark: without calibrating σ_z, σ_ϕ, and ν, the MCMC chain for the MLE will not mix. Given that the MLE MCMC chain will not mix without these calibrations, one would hardly expect the Metropolis within Gibbs GMM chain to mix without them. Indeed, our experience confirms this conjecture.

As mentioned in Section 2, the general principles guiding moment selection are to identify as many parameters as possible from the observed data and try to identify the latent variables themselves indirectly from quantities that can be identified from the observed data. The moment conditions (52) – (60) that follow were designed with these principles in mind.

$$g_1 = (w_t - \rho_\lambda w_{t-1})^2 - \sigma_\lambda^2 \tag{52}$$

$$g_2 = w_{t-1}(w_t - \rho_\lambda w_{t-1}) \tag{53}$$

$$g_3 = [w_{t-1} - (1+\nu)y_{t-1}][w_t - (1+\nu)y_t - \rho_\phi(w_{t-1} - (1+\nu)y_{t-1})] \tag{54}$$

$$g_4 = [w_{t-1} - (1+\nu)y_{t-1}](\phi_t - \rho_\phi \phi_{t-1}) \tag{55}$$

$$g_5 = [w_t - (1+\nu)y_t]^2 - \sigma_\phi^2 \tag{56}$$

$$g_6 = w_{t-1}(y_{t-1} + \frac{1}{\beta}\pi_{t-1} - y_t - \pi_t - \rho_z z_{t-1}) \tag{57}$$

$$g_7 = y_{t-1}(y_{t-1} + \frac{1}{\beta}\pi_{t-1} - y_t - \pi_t - \rho_z z_{t-1}) \tag{58}$$

$$g_8 = \pi_{t-1}(y_{t-1} + \frac{1}{\beta}\pi_{t-1} - y_t - \pi_t - \rho_z z_{t-1}) \tag{59}$$

$$g_9 = (y_{t-1} + \frac{1}{\beta}\pi_{t-1} - y_t - \pi_t)^2 - \frac{\rho_z^2 \sigma_z^2}{1 - \rho_z^2} \tag{60}$$

Conditions (52) and (53) identify ρ_λ and σ_λ. Recalling that ν is calibrated, (54) identifies ρ_ϕ; (55) identifies ϕ_t given ρ_ϕ. (This is not literally true because ϕ_t and ρ_ϕ will interact in the

Metropolis iterations; this qualification applies a few times below also.) Because both ν and σ_ϕ are calibrated, (56) helps enforce an identity linking w_t and y_t. Because σ_z is calibrated, (57) – (59) identify ρ_z, β, and z_t; here we cannot identify ρ_z and β without making use of the latent variable z_t, which is likely to negatively affect GMM relative to MLE. However, (60) does help identify ρ_z and β without using z_t.

One could attempt a comparison with the methods proposed in (Fernandez-Villaverde and Rubio-Ramirez, 2006) using equations (51) to avoid numerical solution methods. The difficulty is that (51) is a singular set of measurement equations, to use the filtering vernacular. The customary approach is to add measurement error to these equations. This presents the additional difficulty of determining how to calibrate the scale of the measurement error. The scale can be manipulated to make results nearly the same as for the MLE (larger scale) or very poor (smaller scale). We do not present these results because we feel one learns nothing from them. One of the advantages of GMM, SMM, and EMM type methods is that singular measurement equations do not cause problems.

Applying the proposed Metropolis within Gibbs GMM method both with and without a Jacobian term to the DSGE model of Subsection 6.2, we obtain the estimates of θ shown in Table 3. Table 3 suggests that the Metropolis within Gibbs GMM estimates are reasonable relative to MLE estimates and within the range one might expect for GMM estimates.

(Table 3 about here)

As mentioned in Section 2, while the moment conditions (52) through (60) can be expected to obtain reasonable results for estimating the parameters θ, they can be expected to do a poor job of estimating the latent variables Λ. That this is the case here can be verified by inspecting figures similar to Figures 7 through 10 that are not shown. In particular, the plots not shown have slopes that are much shallower than those of Figure 8 and 10.

In order to improve the estimate of Λ given X we consider the following additional moment conditions derived from the first order conditions of the DSGE model:

$$h_1 = y_{t-1} + \frac{1}{\beta}\pi_{t-1} - y_t - \pi_t - \rho_z z_{t-1} \tag{61}$$

$$h_2 = w_{t-1}\, h_1 \tag{62}$$

$$h_3 = y_{t-1}\, h_1 \tag{63}$$

$$h_4 = \pi_{t-1}\, h_1 \tag{64}$$

$$h_5 = w_t - (1+\nu)y_t - \phi_t \tag{65}$$

$$h_6 = w_{t-1}\, h_5 \tag{66}$$

$$h_7 = y_{t-1}\, h_5 \tag{67}$$

$$h_8 = \pi_{t-1}\, h_5 \tag{68}$$

Applying the particle filter using conditions (61) through (68) at the true value of θ and $N = 10000$, we obtain the estimates of Λ given X shown as time series plots in Figures 7 and 9, with and without a Jacobian term, respectively, and as scatter plots in Figures 8 and 10, with and without a Jacobian term, respectively.

(Figure 7 about here)

(Figure 8 about here)

(Figure 9 about here)

(Figure 10 about here)

Estimation results using moment conditions (52) through (60) at the Metropolis step and conditions (61) through (68) at the Gibbs step are shown in Table 4. As seen, by comparing Table 3 to Table 4, estimation performance only improves marginally.

(Table 4 about here)

Using moment conditions (52) through (60) at the Metropolis step and conditions (61) through (68) at the Gibbs step rather than conditions (52) through (60) for both reduces computational cost slightly because runtimes for the Gibbs step increase at approximately $RM(T!)N$ whereas runtimes for the Metropolis step increase at approximately $RMTK$.

6.3 Discussion of Examples

The main conclusions from these examples are not surprising, one could have guessed many of them ahead of time:

- In a state space model situation where an analytic form for the measurement equation is available, maximum likelihood when possible, or Flury and Shephard (2010) when not, are better than what we propose unless one is incredibly clever at choosing moment equations.

- When there is no alternative that does not rely on perturbation or numerical approximations that one would rather avoid, our proposal is a viable option.

- The quality of the moment equations matters and there are some principles guiding selection, bearing in mind that the moment equations for the Metropolis step can differ from the moment equations for the Gibbs step:

  - For the Metropolis algorithm to estimate the parameters θ accurately
    - One should identify as many parameters as possible from the observed data.
    - One should make the latent variables depend as much as possible on quantities that can be computed from the observed data.
  - For the Gibbs algorithm to track the trajectory of the unobserved latent variables Λ accurately
    - One should choose moments for the particle filter so that observed variables depend on the latent variables as much as possible without regard for identification of parameters.

- A penalty function can make $p(X, \Lambda, \theta)$ more peaked and improve performance of the particle filter as seen by comparing Figure 1 to Figure 3 and Figure 7 to Figure 9: those with a penalty function have much smaller standard errors. The penalty function we used amounts to letting $p(X, \Lambda, \theta)$ correspond to the distribution of g_T rather than Z.

- A penalty function has little effect on estimates of θ as seen from Tables 2, 3, and 4.

- Bayesian methods are popular for the examples we present because, as seen from the examples, there is not as much information in the data as one could desire. Because the data are calendar dated in most applications, more data is not available. If one

does wish to use moment based Bayesian inference, we conjecture that the technology that we propose here is superior to that proposed by Gallant and Hong (2007).

7 Conclusion

We proposed an algorithm for estimating the parameters of a dynamic model with unobserved variables using only moment conditions and illustrated with two examples: a stochastic volatility model and a dynamic stochastic general equilibrium model. We used a probability distribution derived from a continuously updated GMM criterion considered both with and without a Jacobian term. We found that the Jacobian term had little effect on parameter estimates but did affect the particle filter used in connection with the estimator. Particles deplete much faster when the Jacobian term is present than they do when it is not. (The rate of depletion is the rate at which particle variability declines as t moves from T to 1, compare Figures 1 and 3.) Of interest in applications would be the ability to use our particle filter results to generate impulse response functions for dynamic models with unobserved variables at a given θ using only moment conditions. We have managed to convince ourselves that our results are sufficient for this purpose and are currently working on the requisite algorithms.

8 References

Andrews, Donald W. K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica* 59, 817–858.

Andrieu, C., A. Doucet, and R. Holenstein (2010), "Particle Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society, Series B* 72, 269–342.

Benigno, G., Benigno, P., and Nisticó, S. (2012), "Risk, Monetary Policy and the Exchange Rate," *NBER Macroeconomics Annual* 26, 247–309.

Billingsley, Patrick, and Flemming Topsoe (1967), "Uniformity in Weak Convergence," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 7, 1–16.

Bloom, Nicholas (2009), "The Impact of Uncertainty Shocks," *Econometrica* 77, 623–685.

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

Del Negro, Marco, and Frank Schorfheide (2008), "Forming Priors for DSGE Models (and How it Affects the Assessment of Nominal Rigidities)," *Journal of Monetary Economics* 55, 1191–1208.

Duffie, D. and K. J. Singleton (1993), "Simulated Moments Estimation of Markov Models of Asset Prices," *Econometrica* 61, 929–952.

Fernandez-Villaverde, J., and J. F. Rubio-Ramirez (2006), "Estimating Macroeconomics Models: A Likelihood Approach," NBER Technical Working Paper No. 321.

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

Flury, Thomas, and Neil Shephard (2010), "Bayesian Inference Based Only on Simulated Likelihood: Particle Filter Analysis of Dynamic Economic Models," *Econometric Theory* 27, 933–956.

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.

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

Gallant, A. Ronald, and Halbert L. White, Jr. (1987), *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Basil Blackwell Ltd., Oxford,

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

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica* 50, 1029–1054.

Hansen, L. P., and Singleton, K. J. (1982), "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica* 50 1269–1286.

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

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

The sets that cannot occur when it is known that $\Lambda = \lambda$ are those that do not contain (x, λ) for any x. The conditional probability function $P(X = x \mid \Lambda = \lambda)$ assigns zero probability to the sets that cannot occur. The conditional probability of a set that can occur is computed by dividing its unconditional probability by the probability of the union O_λ of all sets that can occur.

Table 2. Parameter Estimates for the SV Model Moment Conditions (43) through (48) at both the Metropolis and Gibbs Steps.

Parameter	True Value	Mean	Mode	Standard Error
		With Jacobian Term		
ρ	0.25	0.30488	0.30961	0.074778
ϕ	0.8	0.09153	0.94851	0.660790
σ	0.1	0.09023	0.06702	0.050229
		Without Jacobian		
ρ	0.25	0.30271	0.30939	0.076758
ϕ	0.8	0.15348	0.85765	0.643400
σ	0.1	0.11400	0.08435	0.070081
		Flury and Shephard Estimator		
ρ	0.25	0.30278	0.28555	0.059320
ϕ	0.8	0.17599	0.89189	0.509780
σ	0.1	0.09737	0.07839	0.064661

Data of length $T = 250$ was generated by simulating the model of Subsection 6.1 at the parameter values shown in the column labeled "True Value". In the first two panels the model was estimated by using the Metropolis within Gibbs methods described in Section 2 with a one-lag HAC weighting matrix using $N = 1000$ particles for Gibbs and $K = 50$ draws for Metropolis. In the third panel the estimator is the Bayesian estimator proposed by Flury and Shepard (2010) with a flat prior. It is a standard maximum likelihood particle filter estimator except that the seed changes every time a new θ is proposed with N increased as necessary to control the rejection rate of the MCMC chain. The columns labeled mean, mode, and standard deviation are the mean, mode, and standard deviations of a Metropolis within Gibbs chain of length $R = 9637$ for the first two panels and the same from an MCMC chain of length $R = 500000$ with a stride of 5 for the third.

Table 3. Parameter Estimates for the DSGE Model Using Moment Conditions (52) through (60) at Both the Metropolis and Gibbs Steps.

Parameter	True Value	Mean	Mode	Standard Error
		With Jacobian		
ρ_z	0.15	0.21596	0.15006	0.08632
ρ_ϕ	0.68	0.60098	0.58945	0.04988
ρ_λ	0.56	0.50134	0.46443	0.28818
σ_λ	0.11	0.10827	0.08923	0.06494
β	0.996	0.98429	0.99603	0.01476
		Without Jacobian		
ρ_z	0.15	0.21887	0.23069	0.09179
ρ_ϕ	0.68	0.59967	0.60750	0.04988
ρ_λ	0.56	0.50884	0.31473	0.28981
σ_λ	0.11	0.10797	0.11613	0.06896
β	0.996	0.98201	0.99634	0.01834
		Maximum Likelihood		
ρ_z	0.15	0.15165	0.15087	0.00583
ρ_ϕ	0.68	0.59185	0.59419	0.05044
ρ_λ	0.56	0.56207	0.56549	0.05229
σ_λ	0.11	0.11225	0.11189	0.00508
β	0.996	0.99640	0.99643	0.00186

Data of length $T = 250$ was generated by simulating the model of Subsection 6.2 at the parameter values shown in the column labeled "True Value". In the first two panels the model was estimated by using the Metropolis within Gibbs method described in Section 2 with a two-lag HAC weighting matrix using $N = 1000$ particles for Gibbs and $K = 50$ draws for Metropolis. In the third panel the model was estimated by maximum likelihood. The columns labeled mean, mode, and standard deviation are the mean, mode, and standard deviations of a Metropolis within Gibbs chain of length $R = 9637$ for the first two panels and the same from an MCMC chain of length $R = 500000$ with a stride of 5 for the third.

Table 4. Parameter Estimates for the DSGE Model Using Conditions (52) through (60) at the Metropolis Step and Conditions (61) through (68) at the Gibbs Step

Parameter	True Value	Mean	Mode	Standard Error
		With Jacobian		
ρ_z	0.15	0.21702	0.15006	0.08367
ρ_ϕ	0.68	0.61408	0.58945	0.05102
ρ_λ	0.56	0.50082	0.46443	0.28344
σ_λ	0.11	0.11086	0.08924	0.06493
β	0.996	0.98740	0.99603	0.01056
		Without Jacobian		
ρ_z	0.15	0.23508	0.15007	0.08975
ρ_ϕ	0.68	0.69870	0.58945	0.06127
ρ_λ	0.56	0.49904	0.46443	0.28418
σ_λ	0.11	0.11292	0.08924	0.06559
β	0.996	0.97465	0.99604	0.02479
		Maximum Likelihood		
ρ_z	0.15	0.15165	0.15087	0.00583
ρ_ϕ	0.68	0.59185	0.59419	0.05044
ρ_λ	0.56	0.56207	0.56549	0.05229
σ_λ	0.11	0.11225	0.11189	0.00508
β	0.996	0.99640	0.99643	0.00186

As for Table 3.

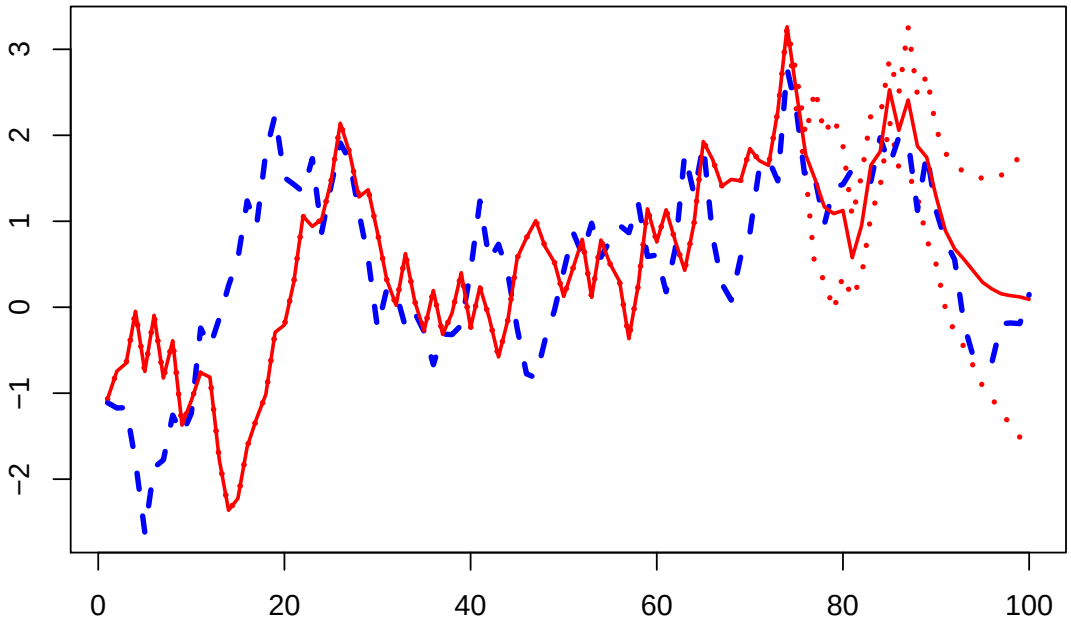

Figure 1. PF for Λ with Jacobian, Time Series Plot, SV Model. Data of length $T = 100$ was generated from a simulation of the model of Subsection 6.1 and $N = 5000$ particles computed using the algorithm described in Section 4.1 with a Jacobian term. The dashed blue line plots the simulated Λ. The solid red line is the mean of the particles and the dotted red lines are plus and minus two pointwise standard errors. The moment equations were (43) through (48); a one lag HAC estimator was used for (4).

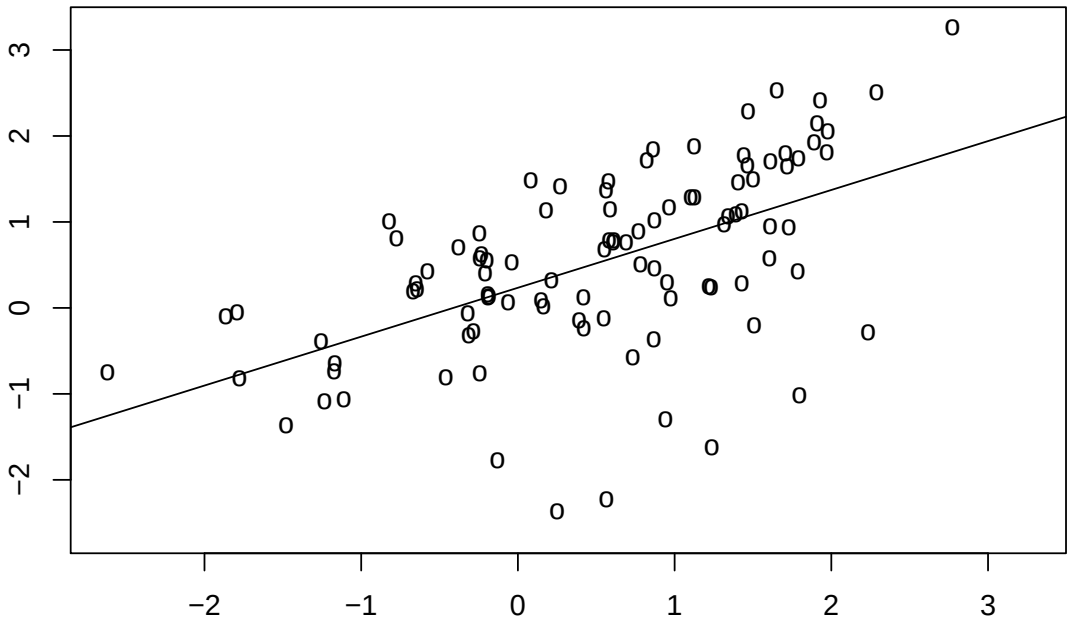

Figure 2. PF for Λ with Jacobian, Scatter Plot, SV Model. As for Figure 1 except that plotted is the mean of the particles vs. the simulated Λ.

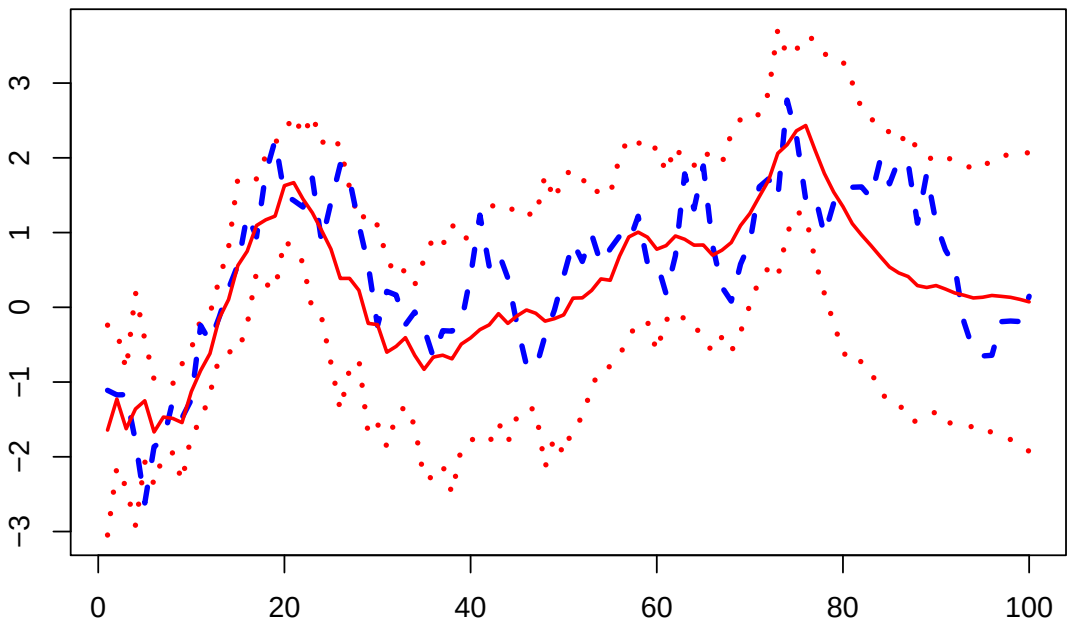

Figure 3. PF for Λ, without Jacobian, Time Series Plot. As for Figure 1 except that estimation is without a Jacobian term.

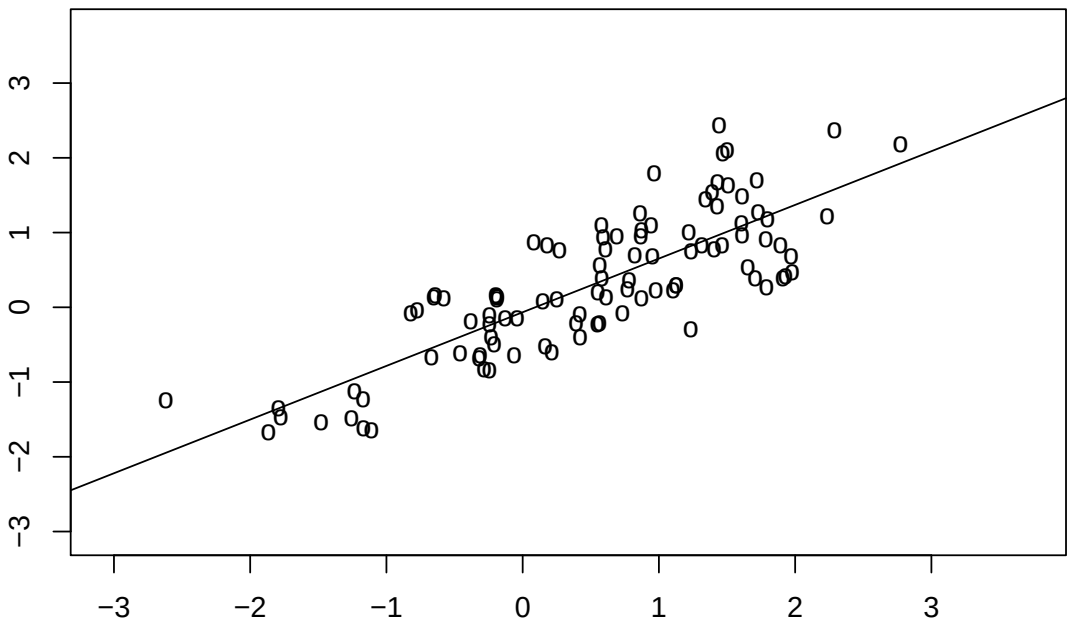

Figure 4. PF for Λ, without Jacobian, Scatter Plot, SV Model. As for Figure 3 except that plotted is the mean of the particles vs. the simulated Λ.

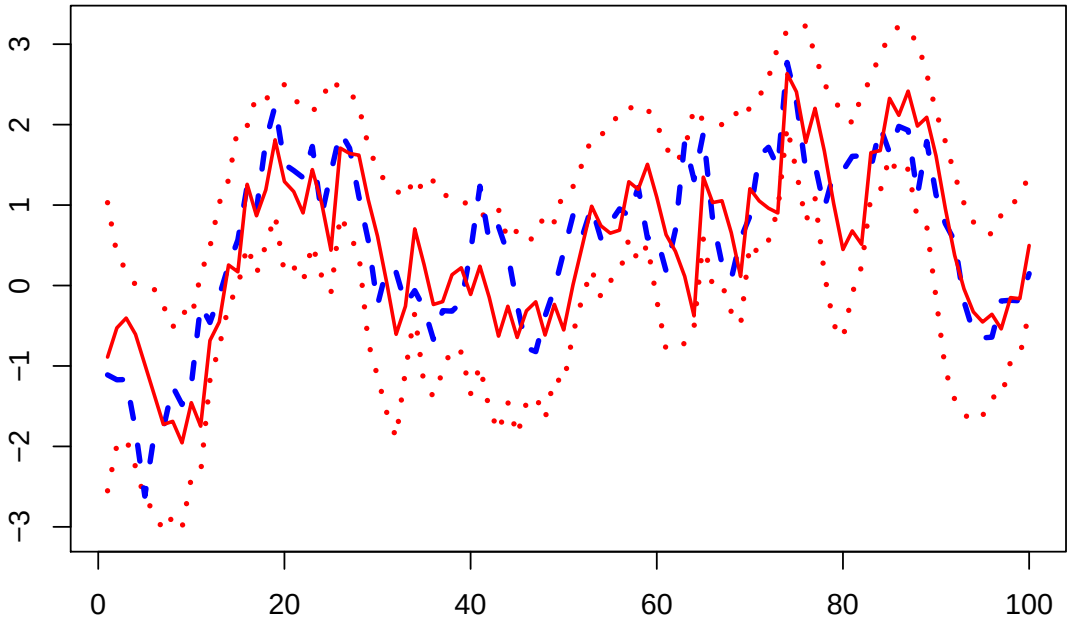

Figure 5. PE for Λ, Flurry-Shephard Method, Time Series Plot, SV Model. As for Figure 1 except that plotted is a filter, not a smooth, and weighting is by the measurement density, not GMM.

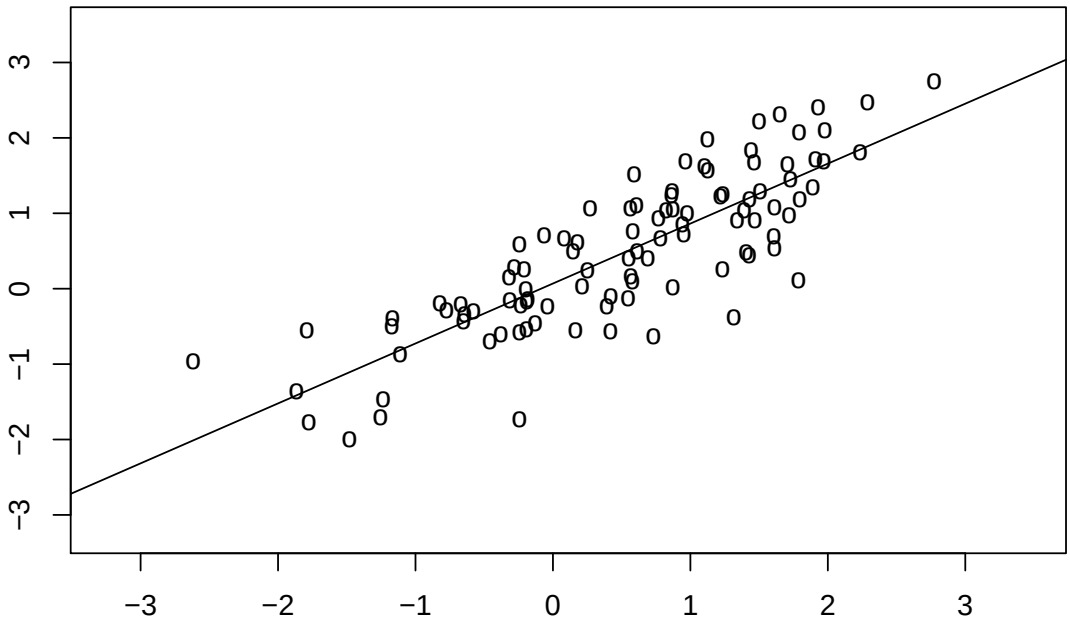

Figure 6. PF for Λ, Flurry-Shephard Method, Scatter Plot. As for Figure 5 except that plotted is the mean of the particles vs. the simulated Λ.

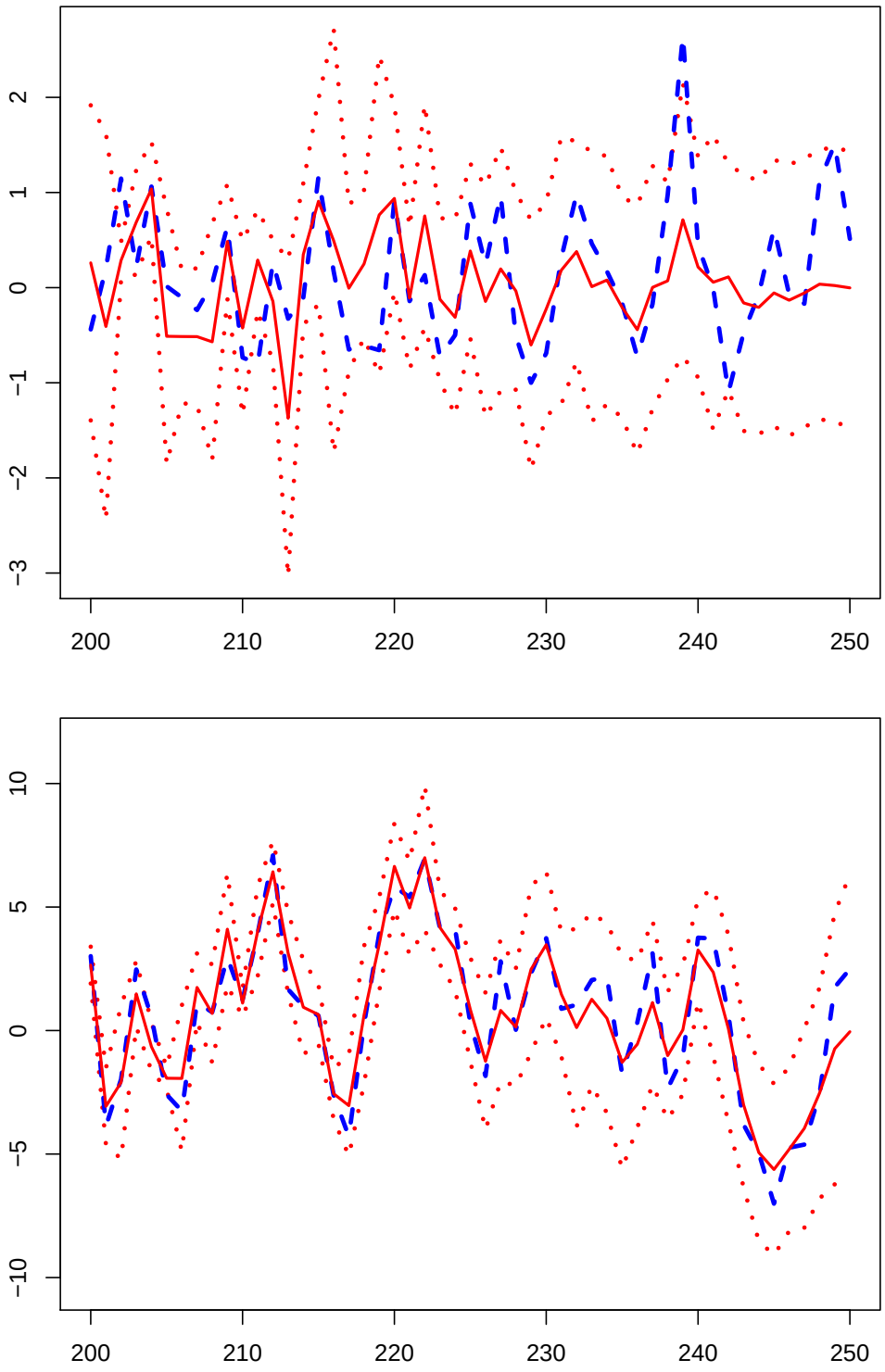

Figure 7. PF for Λ with Jacobian, Time Series Plot, DSGE Model. Data of length $T = 250$ was generated by simulating the model of Subsection 6.2 and $N = 10000$ particles were computed using the algorithm described in Section 4.1 with a Jacobian term. The dashed blue line in the upper panel plots the simulated ϕ_t for the last 50 time points. The lower panel is the same for z_t. In both panels, the solid red line is the mean of the particles and the dotted red lines are plus and minus two pointwise standard errors. The moment equations were (61) through (68); a two lag HAC estimator was used for (4).

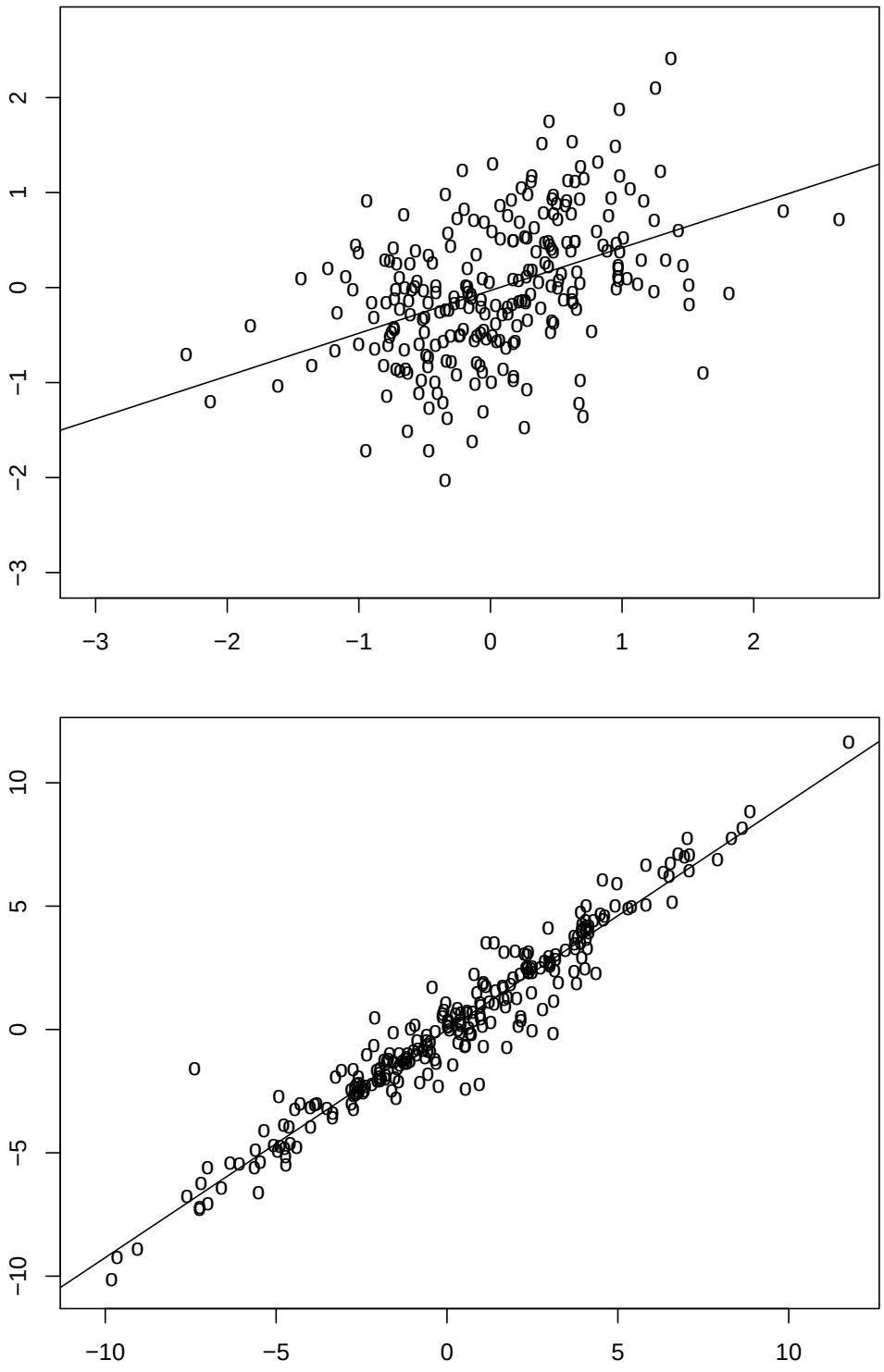

Figure 8. PF for Λ with Jacobian, Scatter Plot, DSGE Model. As for Figure 7 except that plotted is the mean of the particles vs. the simulated Λ for all 250 time points.

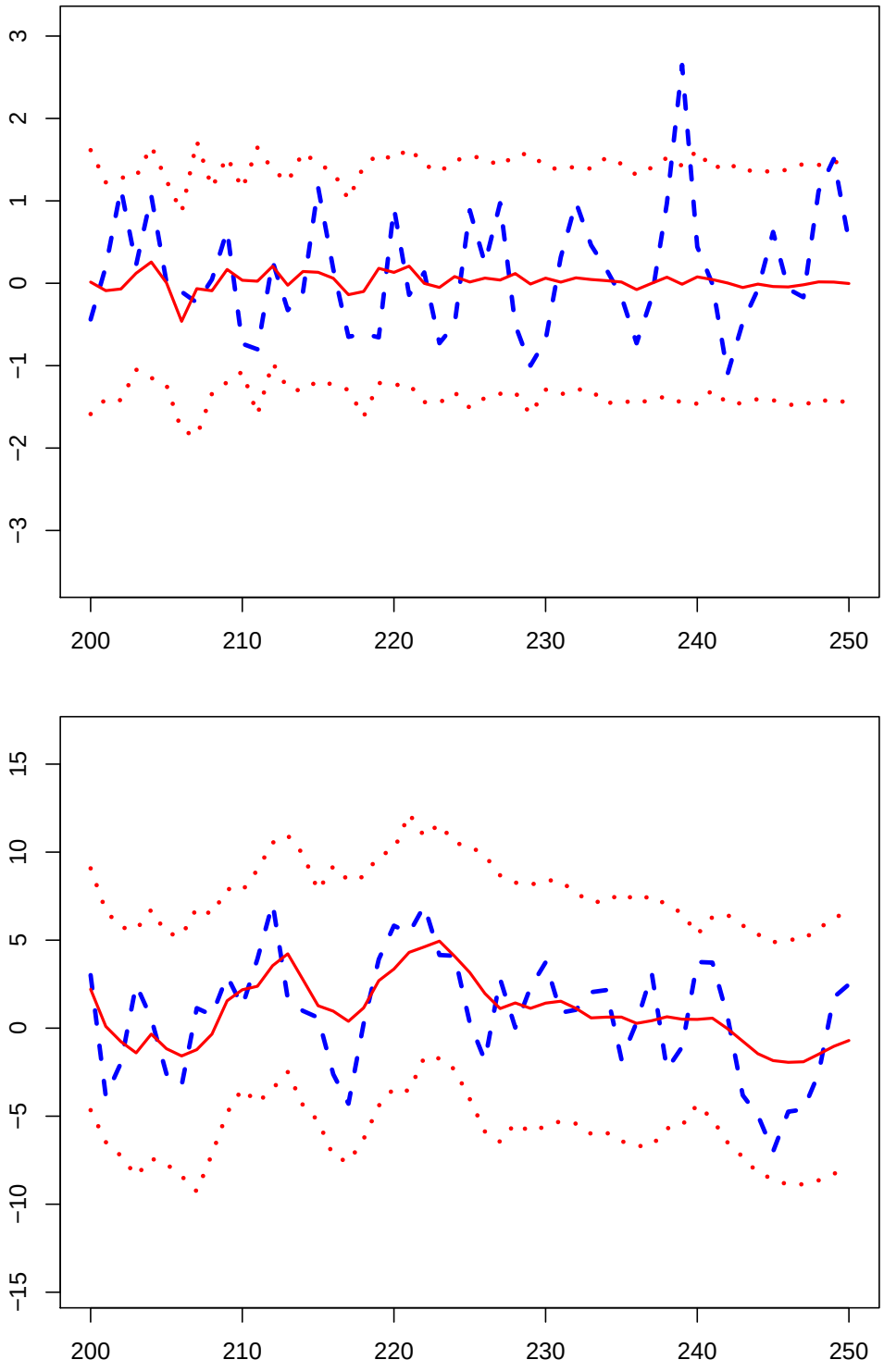

Figure 9. PF for Λ without Jacobian, Time Series Plot, DSGE Model. As for Figure 7 but without a Jacobian term.

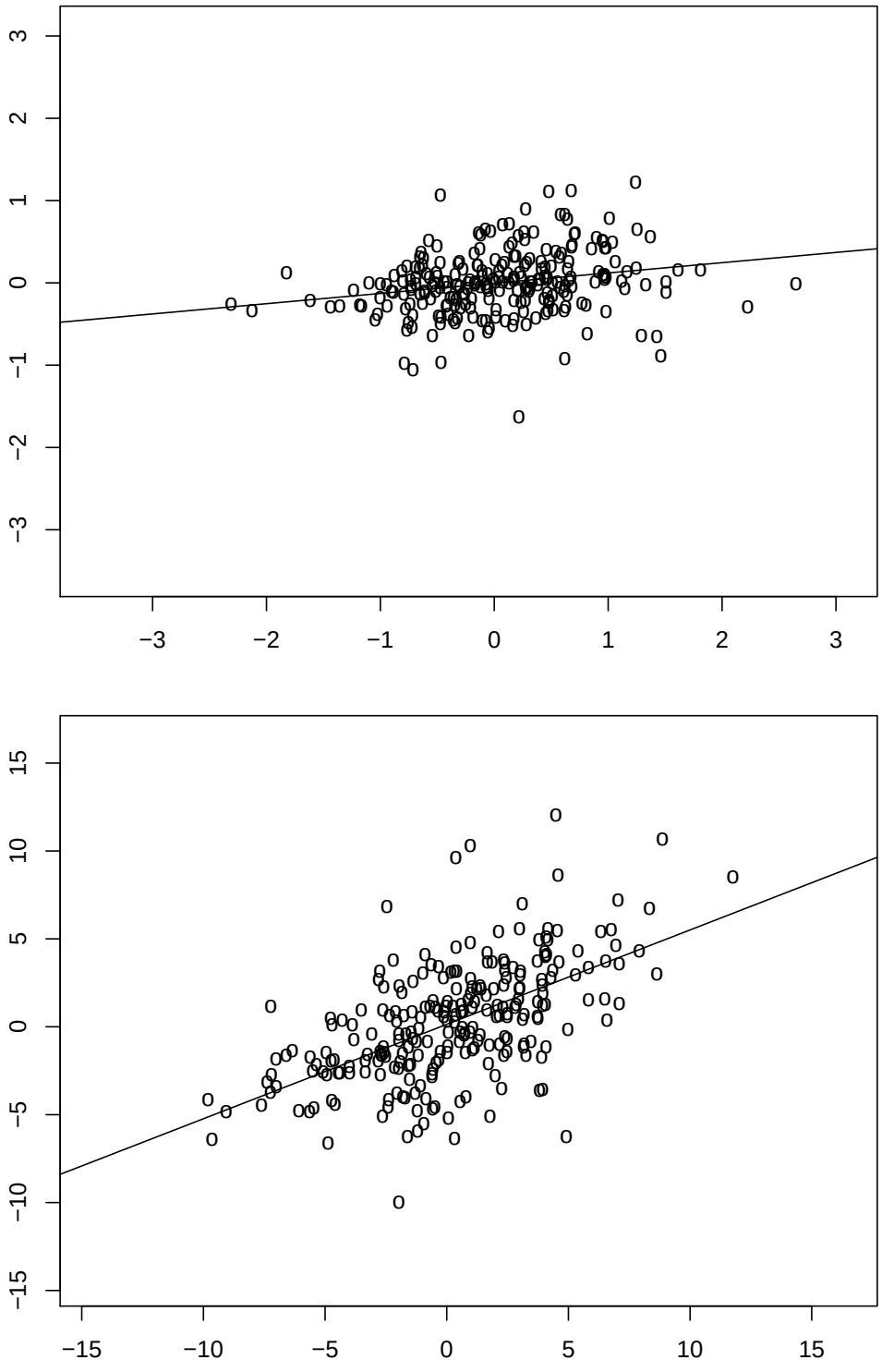

Figure 10. PF for Λ without Jacobian, Scatter Plot, DSGE Model. As for Figure 8 but without a Jacobian term.