cemmap
centre for microdata methods and practice

Generalized method of moments with latent variables

A. Ronald Gallant
Raffaella Giacomini
Giuseppe Ragusa

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP50/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

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: October 5, 2013

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

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 data and must find some means to construct a draw that includes the missing data. We use a distribution that would be correct if the unconditional moment conditions of a GMM criterion defined over complete data were normally distributed to derive a particle filter from which to draw the missing data. This distribution can only assign mass to a limited class of sets. Working around this limitation is the first technical challenge we face. The second is to show that draws based on a normality assumption well approximate draws from the actual finite sample distribution of the moment conditions. What is done here is distinct from the use of a particle filter to integrate out latent variables at some stage of the computation of an estimator. Instead we use the particle filter to construct a proposal density for an MCMC chain.

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 Chernozukov and Hong's (2003) Markov Chain Monte Carlo (MCMC) algorithm as the Metropolis step and Andrieu, Douced, and Holenstein's (2010, Subsection 4.1) modified particle filter algorithm as the Gibbs step. Justification of the Gibbs step is where the aforementioned technical problems arise. 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 GMM 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.

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). 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 stochastic singularity and with possible multiplicity of solutions (indeterminacy).

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 conveniently extend from Bayesian to frequentist inference. That extension is the goal of this paper. Their constructions of various densities derived from a continuously updated GMM criterion are directly applicable. The essence of these constructions can be traced back to fiducial inference (Fisher, 1930). The main issue is showing that a conditional density so constructed 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 citation.

2 Assumptions

ASSUMPTION 1 We require the existence of (but not complete knowledge of) a dynamic structural model that has parameters θ, a vector. 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 $\Lambda_{1:t} = (\Lambda_1, \Lambda_2, ..., \Lambda_t)$.

ASSUMPTION 2 We assume that we can draw from the transition density of the dynamic latent variables $\Lambda_{t+1} \sim P(\Lambda_{t+1} \,|\, \Lambda_t, \theta)$. 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 $P(\Lambda_{t+1} \,|\, \Lambda_t, \theta)$ implies that we can draw from the stationary density $P(\Lambda_t \,|\, \theta)$ by drawing from $P(\Lambda_{t+1} \,|\, \Lambda_t, \theta)$ 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 $P(\Lambda_{t+1} \,|\, \Lambda_{1:t}, X_{1:t}, \theta)$. 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 $P(\Lambda_{t+1} \,|\, \Lambda_{1:t}, X_{1:t}, \theta)$ recursively, and uses the last such draw 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 and Singleton (1982), Gallant and White (1987), and Chernozukov and Hong (2003), which rules out most unit root type behavior. This may require that a parameter lie in an open 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 $g(\cdot, \cdot, \cdot)$ is M-dimensional. The information set is $\mathcal{I}_t = \{X_{-\infty}, ..., X_t, \, \Lambda_{-\infty}, ..., \Lambda_t\}$. We assume that the unconditional moment conditions

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

would identify θ if both X and Λ were observed.

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 prove that the θ draws are a sample from the asymptotic distribution of the GMM estimator determined by (1) 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 only Λ need be identified; for the Metropolis step only θ need be identified. One 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. 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 estimates of, e.g., the mean of the conditional distribution of the latent variables will not. This is 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.

What we require in Assumption 3 is different than what one might expect from the methods proposed by Fernandez-Villaverde and Rubio-Ramirez (2006). We can translate their approach into our context by presuming that the scores of their likelihood (i.e., the joint density for X and Λ) are our moment conditions, in which case the moment conditions would identify Λ and (at least partially) identify the parameters that appear in their likelihood but would not identify the parameters that are unique to the transition density. The reason for the difference is that we are using a modified particle filter (Subsection 4.2) to generate a Gibbs proposal for a metropolis within Gibbs method whereas Fernandez-Villaverde and Rubio-Ramirez are using a full particle filter (Subsection 4.1) to integrate Λ from the likelihood. This integration step puts the parameters unique to the transition density into the objective function. We are not using the particle filter for integration and therefore must identify the parameters unique to the transition density through moment conditions.

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

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

with weighting matrix

$$\Sigma(X,\Lambda,\theta) = \frac{1}{T}\sum_{t=1}^{T}\widetilde{g}(X_t,\Lambda_t,\theta)'\widetilde{g}(X_t,\Lambda_t,\theta) \tag{2}$$

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

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 (2). If a HAC matrix is used, the residuals used to compute it should be of the form shown as (3).

ASSUMPTION 4 We assume that the sample moment conditions normalized by the weighting matrix are asymptotically normal; i.e.,

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

Regularity conditions such that asymptotic normality obtains are in Hansen and Singleton (1982), Gallant and White (1987), and elsewhere.

Define

$$p(X,\Lambda,\theta) = (2\pi)^{-M/2}\exp\left\{-\frac{1}{2}g_T(X,\Lambda,\theta)'\left[\Sigma(X,\Lambda,\theta)\right]^{-1}g_T(X,\Lambda,\theta)\right\} \tag{4}$$

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

$$p(\theta \mid X,\,\Lambda) \propto p(X,\,\Lambda,\,\theta) \tag{5}$$

is a sample from the asymptotic distribution of the GMM estimator for large T.

With Assumption 5 in place, what we have to do to achieve the goal of this paper is discover a method that generates an MCMC chain for (5). We do this by sampling $\left\{\theta^{(i)},\Lambda^{(i)}\right\}$ from the density

$$p(\theta,\Lambda \mid X) \propto p(X,\Lambda,\theta) \tag{6}$$

using a Metropolis within Gibbs algorithm and discarding the Λ draws.

We impose an additional requirement that is a considerable convenience:

ASSUMPTION 6

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(2\pi)^{-M/2}\exp\left\{-\frac{1}{2}g_T(X,\Lambda,\theta)'\left[\Sigma(X,\Lambda,\theta)\right]^{-1}g_T(X,\Lambda,\theta)\right\}dX_1\cdots dX_T=1, \quad (7)$$

which implies

$$p(X\,|\,\Lambda,\,\theta)=p(X,\,\Lambda,\,\theta). \quad (8)$$

As yet we have not encountered a practical application that violates this condition. Usually all that is required is that each element of g is unbounded with respect to an element of X_t and that the residuals used to compute the weighting matrix are centered as in (3). If the integral in (7) does not integrate to one, but one has a convenient means to compute it, then this requirement can be eliminated by using

$$p^{\#}(X\,|\,\Lambda,\theta)=\frac{\exp\left\{-\frac{1}{2}g_T(X,\Lambda,\theta)'\left[\Sigma(X,\Lambda,\theta)\right]^{-1}g_T(X,\Lambda,\theta)\right\}}{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}g_T(X,\Lambda,\theta)'\left[\Sigma(X,\Lambda,\theta)\right]^{-1}g_T(X,\Lambda,\theta)\right\}dX_1\cdots dX_T} \quad (9)$$

instead of $p(X\,|\,\Lambda,\theta)$ to implement the particle filters in Subsections 4.1 and 4.2.

For the Gibbs step we need an additional technical condition:

ASSUMPTION 7 Let $p(X_{1:t},\Lambda_{1:t},\theta)$ denote (4) computed from a partial history. We assume that

$$\mathcal{E}\left\{\log p[X_{1:t},(\Lambda^o_{1:t-1},\Lambda_t),\theta^o]\,|\,\Lambda^o_{1:t}\right\}$$

has an isolated maximum in Λ_t at Λ^o_t if θ^o and $\Lambda^o_{1:t-1}$ were known.

In practice one can check this condition by looking at a few plots of $\log p(X_{1:t},\Lambda_{1:t},\theta)$ against Λ_t with all else fixed to see if Λ_t appears to be identified. If not, one can change the moment conditions. In this connection see the discussion of moment conditions (41) through (48) in Subsection 6.2.

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 $p(\theta\,|\,X,\Lambda^{(i-1)})$ knowing $\theta^{(i-1)}$ using a Metropolis algorithm (Subsection 4.3).

3. Sample $\Lambda^{(i)}$ from $p(\Lambda \mid X, \theta^{(i)})$ knowing $\Lambda^{(i-1)}$, where

$$p(\Lambda \mid X, \theta) \propto p(X, \Lambda, \theta), \tag{10}$$

   using a modified particle filter (Subsection 4.2).

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

We discard the Λ draws to avoid excessive consumption of disk space and because there is little use for the joint distribution of θ and Λ in frequentist inference. Of more use is 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 $P(\Lambda \mid X,\, \theta)$. Actually, as we shall see in Section 5, it is only the ordinary particle filter algorithm that we have to derive because the rest of the theory follows directly by citation.

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 Section 5. Until Section 5 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 up to that point.

Sometimes one uses a penalty function in connection with MCMC using (6). In our examples we shall investigate the effect of multiplying (6) by a Jacobian term $[\det \Sigma(X, \Lambda, \theta)]^{-M/2}$.

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 restate the subset of their results that we need here in a form more suited to frequentist inference.

3.1 The Simplest Example

(Figure 1 about here)

Consider a random sample $X_1, \ldots, X_n$ from a normal distribution whose mean Λ is also normally distributed. That is, there is one draw to get Λ and then n draws from $n(X \mid \Lambda, \sigma^2)$. The statistic

$$Z = \sqrt{n}\left(\frac{\bar{X} - \Lambda}{\sqrt{s^2}}\right)$$

will have the t-distribution, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. For large enough n the t-distribution cannot be distinguished from the normal, which, for simplicity, is the distribution that we shall use for Z to illustrate the ideas. With this simplification, $\bar{X}$ and Λ have joint density

$$p(\bar{X}, \Lambda) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{n}{2}\left(\frac{\bar{X}-\Lambda}{\sigma}\right)^2}.$$

This assertion is verified by an application of the change of measure formula (Gallant and Hong, 2007).

This density is nonstandard in the sense that joint probability over $(\bar{X}, \Lambda) \in \mathbb{R}^2$ can only be assigned to (the smallest σ-algebra containing all) sets bounded by 45 degree lines. An example is the set labeled $A_{(\bar{X},\Lambda)}$ in Figure 1. The conditional probability for a set such as that labeled $C_{(\bar{X}|\Lambda)}$ in Figure 1 is computed as

$$P(C \mid \Lambda) = \frac{\int_C p(\bar{X}, \Lambda)\, d\bar{X}}{\int_{-\infty}^{\infty} p(\bar{X}, \Lambda)\, d\bar{X}}.$$

Conditional probability must be computed in this way to achieve coherency. In most applications, as in this one, the integral that appears in the denominator of $P(C \mid \Lambda)$ will be identically equal to one for all Λ. Therefore, because the denominator is identically one, $p(\bar{X}, \Lambda)$ is also a conditional density.

The conditional probability $P(C \mid \Lambda)$ also attaches itself to sets of the form $C^n = \{(X_1, \ldots, X_n) : \bar{X} \in C\}$ by the change of measure formula (Gallant and Hong, 2007). Information is lost relative to the full likelihood $p(X_1, \ldots, X_n \mid \Lambda)$, were it available, because only (the σ-algebra containing all) sets of the form C^n in $\mathbb{R}^n$ can be assigned conditional probability by the density $p(\bar{X}, \Lambda)$. Denote the smallest σ-algebra containing all sets of the form C^n by $\mathcal{C}^n$. Nonempty sets C^n in $\mathcal{C}^n$ will be unbounded and have the restriction, among others, that if $(X_1, X_2, \ldots, X_n)$ is in C^n then so will $(X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(n)})$ be in C^n for any permutation $\sigma(\cdot)$ of the integers 1 through n. In particular, bounded rectangles in $\mathbb{R}^n$ will

not be in this σ-algebra and therefore cannot be assigned conditional probability whereas they can be assigned probability by the full likelihood. This implies that $\mathcal{C}^n$ cannot contain all the Borel subsets of $\mathbb{R}^n$.

The essential point of this subsection is that $p(\bar{X}, \Lambda)$ can be regarded as a conditional density on $\mathbb{R}^n$ and is therefore a likelihood.

A corollary is obvious at sight from the change of measure formula but is not emphasized in Gallant and Hong (2007) because it is largely irrelevant to Bayesian inference: If a set C^n is in $\mathcal{C}^n$, then it is assigned the same probability by both the full likelihood and the likelihood induced by the GMM criterion. This means that if one computes the unconditional expectation of $g(X, \Lambda)$ (where $g(\cdot, \Lambda)$ measurable $\mathcal{C}^n$) using the law of iterated expectations by integrating first with respect to X using either of the two likelihoods and then with respect to Λ, one gets the same answer.

3.2 The Abstraction

The abstraction from these ideas is that the density $p(X, \Lambda, \theta)$ given by (4) generates a likelihood $p(X \,|\, \Lambda, \theta) = p(X, \Lambda, \theta)$ when (7) holds. If (7) fails, then one uses $p^{\#}(X \,|\, \Lambda, \theta)$ given by (9) if computationally feasible. As it is obvious where $p^{\#}(X \,|\, \Lambda, \theta)$ needs to be substituted for $p(X \,|\, \Lambda, \theta)$, we shall use $p(X \,|\, \Lambda, \theta)$ and the transition density $p(\Lambda_{t+1} \,|\, \Lambda_t, \theta)$ as the basis for frequentist inference in the sequel without further comment.

We remark in passing that if one conducts Bayesian inference using $p(X \,|\, \Lambda, \theta)$ as a likelihood and $p(\Lambda_{t+1} \,|\, \Lambda_t, \theta) \times p(\theta)$ as a prior for some density $p(\theta)$, as proposed by Gallant and Hong (2007), we expect that using $p(X \,|\, \Lambda, \theta) \times p(\theta)$ in the Metropolis part of the method proposed here and leaving the rest unchanged will prove to be a better algorithm than the one proposed by Gallant and Hong.

4 Algorithms

Three algorithms are required to implement our method:

- A particle filter (PF) algorithm.
  - Input: θ.
  - Output: Draws $\left\{\Lambda^{(i)}\right\}_{i=1}^{R}$ from $P(\Lambda \,|\, X, \theta)$.

- A Gibbs algorithm.
  - Input: The previous draw $\Lambda^{(i-1)}$ and a draw $\theta^{(i)}$ from $p(\theta \mid X, \Lambda^{(i-1)})$.
  - Output: A draw $\Lambda^{(i)}$ from $P(\Lambda \mid X, \theta^{(i)})$.
- 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 $p(\theta \mid X, \Lambda^{(i)})$.

In this section we present them in turn.

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

$$p(X_{1:t}, \Lambda_{1:t}, \theta) = (2\pi)^{-M/2} \exp\left\{-\frac{1}{2} g_t(X_{1:t}, \Lambda_{1:t}, \theta)' \left[\Sigma(X_{1:t}, \Lambda_{1:t}, \theta)\right]^{-1} g_t(X_{1:t}, \Lambda_{1:t}, \theta)\right\}, \quad (11)$$

which corresponds to (4). The densities $p(X_{1:t} \mid \Lambda_{1:t}, \theta)$ and $p(\theta \mid \Lambda_{1:t}, X_{1:t})$ are proportional to (11). For $p(X_{1:t} \mid \Lambda_{1:t}, \theta)$ the proportionality factor is assumed to be one; 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 $p(\Lambda_t | \Lambda_{t-1}, \gamma)$.
- 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 $p(\Lambda_t | \Lambda_{t-1}^{(i)})$ 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 $p(\Lambda_t | \Lambda_{t-1}, \gamma)$.
   - 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 $p(\Lambda_t | \Lambda_{t-1}^{(i)})$ 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 $K = 50$.

- 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 7, the particle filter algorithm defined in Subsection 4.1 generates draws from $P(\Lambda \,|\, X, \theta)$.

Proof Define

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

and

$$Z_T(X, \Lambda, \theta) = [\Sigma(X, \Lambda, \theta)]^{-1/2} g_T(X, \Lambda, \theta).$$

Let θ^o denote the true value of θ and let $\Lambda^o_{1:t}$ and $X^o_{1:t}$ denote the realized values of the data and latent variables. Neither θ^o nor $\Lambda^o_{1:t}$ are observed; $X^o_{1:t}$ is observed. Let $z^o_t = Z_t(X^o_{1:t}, \Lambda^o_{1:t}, \theta^o)$.

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 $B_{(\Lambda_{1:t},\theta)} = \{z : z = Z_t(X_{1:t}, \Lambda_{1:t}, \theta),\, X_{1:t} \in \mathcal{X}_{(\Lambda_{1:t},\theta)}\}$. We have assumed that $\int_{B(\Lambda_{1:t},\theta)} n(z|0, I)\, dz = 1$. Under this assumption, $p(X_{1:t}, \Lambda_{1:t}, \theta)$ can be regarded as a conditional density for $X_{1:t}$ given $\Lambda_{1:t}$ that can assign conditional probability to sets of the form

$$C_{1:t} = \{X_{1:t} \,:\, Z_t(X_{1:t}, \Lambda_{1:t}, \theta) \in B\}$$

where $B \subset \mathbb{R}^M$ is Borel. The probability assigned to $C_{1:t}$ is $P(C_{1:t}|\Lambda_{1:t}, \theta) = \int_B n(z|0, I)\, dz$. In the case B is a singleton, we use the notation $C^z_{1:t}$. Let $\mathcal{C}_{1:t}$ denote the smallest σ-algebra containing the $C_{1:t}$.

The functions $f(\cdot)$ for which the integral $\int f(X_{1:t})\, P(dX_{1:t} \,|\, \Lambda_{1:t}, \theta)$ can be computed must be measurable with respect to $\mathcal{C}_{1:t}$. Such $f(\cdot)$ will be constant on $C^z_{1:t}$.

Given $(\Lambda_{1:t}, \theta)$, for each z choose a point $X^*_{1:t} \in \mathcal{X}_{(\Lambda_{1:t},\theta)}$ for which

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

and set

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

Conversely, any realization $X_{1:t}$ that is possible under the pair $(\Lambda_{1:t}, \theta)$ must lie in some $C^z_{1:t}$ thus giving a map $X_{1:t} \to X^*_{1:t} \to z_t$ in the opposite direction. Note that if $X^*_{1:t} \to z^*_t$ then $z^*_t \to X^*_{1:t}$; therefore, for convenience, we will always choose $X^o_{1:t}$ as the $X^*_{1:t}$ for its image so that $X^o_{1:t} \to z^o_t \to X^o_{1:t}$.

The following two points are subtle but important: (1) With $\Lambda_{1:t}$ and θ held fixed, an $f(\cdot)$ measurable with respect to $\mathcal{C}_{1:t}$ can be regarded either as a function of z_t or as a function

of $X_{1:t}$. (2) A function $g(\cdot)$ of the form

$$g(z_{1:t}) = f[X_{1:1}(z_1, \Lambda_{1:t}, \theta), X_{1:2}(z_2, \Lambda_{1:t}, \theta), \ldots, X_{1:t}(z_t, \Lambda_{1:t}, \theta)] \tag{12}$$

can be evaluated at $(z^o_{1:t}, \Lambda_{1:t}, \theta)$ using

$$g(z^o_{1:t}) = f[X^o_{1:1}, X^o_{1:2}, \ldots, X^o_{1:t}].$$

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

$$p(z_t \,|\, \Lambda_{1:t}, \theta) = n\,\{[Z_t[X_{1:t}(z_t, \Lambda_{1:t}, \theta), \Lambda_{1:t}, \theta] \,|\, 0, I\} \tag{13}$$

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 $\Lambda^{(i)}_{1:T}$ from the density $p(\Lambda_{1:T} \,|\, z_{1:T}, \theta)$.

What we want are draws from the actual conditional density of $\Lambda = \Lambda_{1:T}$ given $X^o_{1:T}$ that we denote by $f_T(\Lambda \,|\, z_{1:T}, \theta)$. Let $\Psi_T(\cdot)$ denote the actual distribution of $Z_T(X^o_{1:T}, \Lambda, \theta)$ 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

$$u^{(i)}_T = \phi(z^{(i)}_T)\, p(\Lambda \,|\, \theta) \tag{14}$$

$$U_T = \int \phi(Z_T(X^o_{1:T}, \Lambda, \theta))\, p(\Lambda \,|\, \theta)\, d\Lambda \tag{15}$$

$$v^{(i)}_T = \psi_T(z^{(i)}_T)\, p(\Lambda \,|\, \theta) \tag{16}$$

$$V_T = \int \psi_T(Z_T(X^o_{1:T}, \Lambda, \theta))\, p(\Lambda \,|\, \theta)\, d\Lambda \tag{17}$$

where

$$p(\Lambda|\theta) = p(\Lambda^{(i)}_1 \,|\, \theta) \prod_{s=2}^{T} p(\Lambda^{(i)}_s \,|\, \Lambda^{(i)}_{s-1}, \theta).$$

Using (14) through (17) to construct importance sampling weights, we have

$$\frac{1}{N}\sum_{i=1}^{N} \frac{v^{(i)}_T}{u^{(i)}_T}\frac{U_T}{V_T}\, g_T(X^o e :_{1:T}, \Lambda^{(i)}_{1:T}, \theta) = \frac{U_T}{V_T}\frac{1}{N}\sum_{i=1}^{N} \frac{\psi_T(z^{(i)}_T)}{\phi(z^{(i)}_T)}\, g_T(X^o_{1:T}, \Lambda^{(i)}_{1:T}, \theta) \tag{18}$$

is an approximation to

$$\int g_T(X^o_{1:T}, \Lambda, \theta)\, f_T(\Lambda \,|\, z_{1:T}, \theta)\, d\Lambda \tag{19}$$

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

We shall first show that

$$\frac{U_T}{V_T}\frac{1}{N}\sum_{i=1}^{N} g_T(X^o_{1:T}, \Lambda^{(i)}_{1:T}, \theta) \tag{20}$$

also approximates (19) for large N and T.

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

$$\frac{U_T}{V_T}\int I\{Z_T(X^o_{1:T}, \Lambda, \theta) \in (a_0, b_0]\}\, g_T(X^o_{1:T}, \Lambda, \theta)\, f_T(\Lambda \,|\, z_{1:T}, \theta)\, d\Lambda \tag{21}$$

approximates (19) 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

$$\frac{U_T}{V_T}\frac{1}{N}\sum_{i=1}^{N} I\{Z_T(X^o_{1:T}, \Lambda, \theta) \in (a_0, b_0]\}\, \frac{\psi_T(z_T^{(i)})}{\phi(z_T^{(i)})}\, g_T(X^o_{1:T}, \Lambda^{(i)}_{1:T}, \theta) \tag{22}$$

approximates (21) 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 (20) approximates (19) 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 $1 = \sum_{j=1}^{J} \int I_{I_j}(Z_T(X^o_{1:T}, \Lambda, \theta))\, p(\Lambda|\theta)\, d\Lambda$. Then for any T,

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

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

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

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

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

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, Douced, 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.

The results of Chernozukov and Hong (2003) require that (1) hold, at least in the limit as $T \to \infty$. As noted in the proof of Theorem 1, the change of measure formula implies that the expectation is the same for functions measurable $\mathcal{C}_{1:T}$ whether one computes it by first computing the conditional expectation with respect to Λ given X using the full likelihood and then integrating with respect to X or by using the density $f_T(\Lambda \,|\, z_{1:T}, \theta)$ to compute the conditional expectation. As shown in Subsection 5.1, the error in computing an expectations using a draw from $p(\Lambda \,|\, X, \theta)$ instead of $f_T(\Lambda \,|\, z_{1:T}, \theta)$ tends to zero as $T \to \infty$. An implication is that the difference between the approximation of $\mathcal{E}[g_T(X, \Lambda, \Theta)]$ gotten by using a draw (Λ, X) from the full likelihood, i.e., data, and gotten by first drawing X (i.e., data), then drawing Λ from $f_T(\Lambda \,|\, z_{1:T}, \theta)$ tends to zero with T.

5.2 Gibbs Theory

The proof above that we can draw a sample from $f_T(\Lambda \,|\, z_{1:T}, \theta)$ with negligible error for large T implies that the algorithm given in Subsection 4.1 of Andrieu, Douced, 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 7.

5.2.1 Comments on Gibbs Theory

Using only one particle to evaluate the conditional expectation does seem wasteful when N are available from the modified particle filter of Subsection 4.1, as does carrying only one particle forward. When we modified our code to carry all particles forward, to draw N new particles from the $2N$ old and new particles at each selection step, and to average the moments and weighting matrix over all particles before computing Z, we found that the only

effect was to change results slightly at the cost of increasing run times by a factor of about $2(T!)N$. We therefore dismissed retaining and using more than one particle from further consideration.

In our examples we found that $N = 1000$ gave about the same results as $N = 5000$ and larger. Andrieu, Douced, 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 the Metropolis within Gibbs algorithm will mix (Gamerman and Lopes, 2006).

6 Examples

We illustrate our proposal with two examples.

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:

$$
\begin{aligned}
g_1 &= (X_t - \rho X_{t-1})^2 - [\exp(\Lambda_t)]^2 && (23)\\
g_2 &= |X_t - \rho X_{t-1}||X_{t-1} - \rho X_{t-2}| - \left(\frac{2}{\pi}\right)^2 \exp(\Lambda_t)\exp(\Lambda_{t-1}) && (24)\\
&\vdots \\
g_{L+1} &= |X_t - \rho X_{t-1}||X_{t-L} - \rho X_{t-L-1}| - \left(\frac{2}{\pi}\right)^2 \exp(\Lambda_t)\exp(\Lambda_{t-L}) && (25)\\
g_{L+2} &= X_{t-1}(X_t - \rho X_{t-1}) && (26)\\
g_{L+3} &= \Lambda_{t-1}(\Lambda_t - \phi\Lambda_{t-1}) && (27)\\
g_{L+4} &= (\Lambda_t - \phi\Lambda_{t-1})^2 - \sigma^2 && (28)
\end{aligned}
$$

Moment (26) identifies ρ independently of Λ_t; moments (23) through (26) overidentify Λ_t given ρ. Moment (27) identifies ϕ given Λ_t and moment (28) identifies σ given Λ_t and ϕ.

What may not be obvious here is how an equation such as (23) 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 (23) 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.

Estimates of θ for the SV model are shown in Table 1 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 2 and as a scatter plot in Figure 3 for the case when a Jacobian term is included and as Figures 4 and 5 when it is not. The plots for the

Flury and Shephard estimator are Figures 6 and 7. 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 1 about here)

(Figure 2 about here)

(Figure 3 about here)

(Figure 4 about here)

(Figure 5 about here)

(Figure 6 about here)

(Figure 7 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

$$\begin{aligned} z_t &= \rho_z z_{t-1} + \sigma_z \epsilon_{z,t} \\ \phi_t &= \rho_\phi \phi_{t-1} + \sigma_\phi \epsilon_{\phi,t} \\ \lambda_t &= \rho_\lambda \lambda_{t-1} + \sigma_\lambda \epsilon_{\lambda,t}, \end{aligned} \tag{29}$$

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

The first order conditions are

$$\begin{aligned} 0 &= y_t + \frac{1}{\beta}\pi_t - \mathcal{E}_t(y_{t+1} + \pi_{t+1} + z_{t+1}) \\ 0 &= w_t + \lambda_t \\ 0 &= w_t - (1+\nu) y_t - \phi_t \end{aligned} \tag{30}$$

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

The solution for the endogenous variables is

$$\begin{aligned} w_t &= -\lambda_t \\ y_t &= -\frac{1}{1+\nu}\lambda_t - \frac{1}{1+\nu}\phi_t \\ \pi_t &= \beta \frac{1-\rho_\lambda}{(1+\nu)(1-\beta\rho_\lambda)}\lambda_t + \beta \frac{1-\rho_\phi}{(1+\nu)(1-\beta\rho_\phi)}\phi_t + \beta \frac{\rho_z}{(1-\beta\rho_z)} z_t \end{aligned} \tag{31}$$

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 (29) into equations (31). 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 (Chernozukov 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.

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 (32) – (40) that follow were designed with these principles in mind.

$$g_1 = (w_t - \rho_\lambda w_{t-1})^2 - \sigma_\lambda^2 \tag{32}$$

$$g_2 = w_{t-1}(w_t - \rho_\lambda w_{t-1}) \tag{33}$$

$$g_3 = [w_{t-1} - (1+\nu)y_{t-1}][w_t - (1+\nu)y_t - \rho_\phi(w_{t-1} - (1+\nu)y_{t-1})] \tag{34}$$

$$g_4 = [w_{t-1} - (1+\nu)y_{t-1}](\phi_t - \rho_\phi \phi_{t-1}) \tag{35}$$

$$g_5 = [w_t - (1+\nu)y_t]^2 - \sigma_\phi^2 \tag{36}$$

$$g_6 = w_{t-1}(y_{t-1} + \frac{1}{\beta}\pi_{t-1} - y_t - \pi_t - \rho_z z_{t-1}) \tag{37}$$

$$g_7 = y_{t-1}(y_{t-1} + \frac{1}{\beta}\pi_{t-1} - y_t - \pi_t - \rho_z z_{t-1}) \tag{38}$$

$$g_8 = \pi_{t-1}(y_{t-1} + \frac{1}{\beta}\pi_{t-1} - y_t - \pi_t - \rho_z z_{t-1}) \tag{39}$$

$$g_9 = (y_{t-1} + \frac{1}{\beta}\pi_{t-1} - y_t - \pi_t)^2 - \frac{\rho_z^2 \sigma_z^2}{1 - \rho_z^2} \tag{40}$$

Conditions (32) and (33) identify ρ_λ and σ_λ. Recalling that ν is calibrated, (34) identifies ρ_ϕ; (35) 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, (36) helps enforce an identity linking w_t and y_t. Because σ_z is calibrated, (37) – (39) 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, (40) 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 (31) to avoid numerical solution methods. The difficulty is that (31) 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 2. Table 2 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 2 about here)

As mentioned in Section 2, while the moment conditions (32) through (40) 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 as can be verified by inspecting figures similar to Figures 8 through 11 that are not shown. In particular, the plots not shown have slopes that are much shallower than those of Figure 9 and 11.

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:

$$h_1 = y_{t-1} + \frac{1}{\beta}\pi_{t-1} - y_t - \pi_t - \rho_z z_{t-1} \tag{41}$$

$$h_2 = w_{t-1}\, h_1 \tag{42}$$

$$h_3 = y_{t-1}\, h_1 \tag{43}$$

$$h_4 = \pi_{t-1}\, h_1 \tag{44}$$

$$h_5 = w_t - (1+\nu)y_t - \phi_t \tag{45}$$

$$h_6 = w_{t-1}\, h_5 \tag{46}$$

$$h_7 = y_{t-1}\, h_5 \tag{47}$$

$$h_8 = \pi_{t-1}\, h_5 \tag{48}$$

Applying the particle filter using conditions (41) through (48) at the true value of θ and $N = 10000$, we obtain the estimates of Λ given X shown as time series plots in Figures 8 and 10, with and without a Jacobian term, respectively, and as scatter plots in Figures 9 and 11, with and without a Jacobian term, respectively.

(Figure 8 about here)

(Figure 9 about here)

(Figure 10 about here)

(Figure 11 about here)

Estimation results using moment conditions (32) through (40) at the Metropolis step and conditions (41) through (48) at the Gibbs step are shown in Table 3. As seen, by comparing Table 2 to Table 3, estimation performance only improves marginally.

(Table 3 about here)

Using moment conditions (32) through (40) at the Metropolis step and conditions (41) through (48) at the Gibbs step rather than conditions (32) through (40) 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 $RMTK$.

7 Discussion of Examples

The main conclusions from these results are not surprising, one could have guessed most 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 chosen moments does matter and there are some principles guiding selection in this context:
  - One should identify as many parameters as possible from the observed data.
  - One should identify the latent variables themselves from quantities that can be identified from the observed data.
- A penalty function can make $p(X, \Lambda, \theta)$ more peaked and improve performance as seen most dramatically by comparing the figures for PFs computed with and without a Jacobian: those with 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_T.
- Bayesian methods are popular for the examples we present because, as seen from the examples, there is not at all as much information in the data as one could desire. And, because the data are calendar dated in most applications, more is not available. If one does wish to use moment based Bayesian inference, we conjecture that the technology that we propose is superior to that proposed by Gallant and Hong (2007).

8 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 stochas-

tic volatility model and a dynamic stochastic general equilibrium model. We used both a continuously updated GMM criterion and the same with a Jacobian penalty term. We found that estimates improved slightly with the Jacobian term. 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. E.g., compare Figures 2 and 4.) More relevant to 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.

9 References

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

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

Billingsley, Patrick, and Flemming Topsoe (1967), "Uniformity in Weak Convergence," *Z. Wahrscheinlichkeitstheorie verw. Geb.* 7, 1–16.

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.

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*, forthcoming.

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

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,* forthcoming.

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

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

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

Table 1. Parameter Estimates for the SV Model Moment Conditions (23) through (28) 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 2. Parameter Estimates for the DSGE Model Using Moment Conditions (32) through (40) 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 3. Parameter Estimates for the DSGE Model Using Conditions (32) through (40) at the Metropolis Step and Conditions (41) through (48) 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 2.

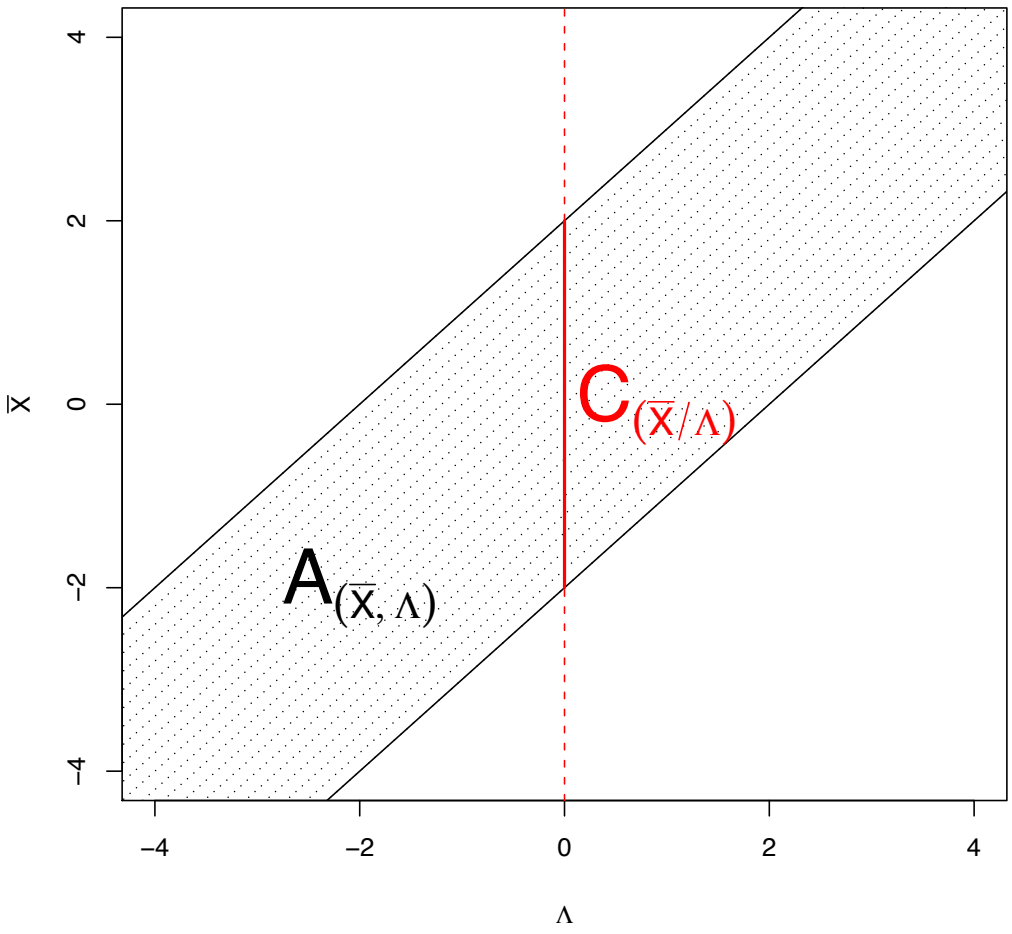

Figure 1. GMM Probability Assignment. Under the assumption that $\bar{X}$ and Λ have joint density $p(\bar{X}, \Lambda) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{n}{2}\frac{(\bar{X}-\Lambda)^2}{\sigma^2}}$ where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, joint probability on $(\bar{X}, \Lambda)$ can only be assigned to sets bounded by 45 degree lines such as the one labeled $A_{(\bar{X},\Lambda)}$ in the figure. The conditional probability for a set such as $C_{(\bar{X}|\Lambda)}$ in the figure is computed as

$$P(C \,|\, \Lambda) = \frac{\int_C p(\bar{X}, \Lambda)\, d\bar{X}}{\int_{-\infty}^{\infty} p(\bar{X}, \Lambda)\, d\bar{X}}$$

The conditional probability $P(C \,|\, \Lambda)$ also attaches itself to sets of the form $C^n = \{(X_1, \ldots, X_n) \,:\, \bar{X} \in C\,\}$ by the change of measure formula. Information is lost relative to the full likelihood $p(X_1, \ldots, X_n \,|\, \Lambda)$ because only the σ-algebra containing all sets of the form C^n in $\mathbb{R}^n$ can be assigned conditional probability by the density $p(\bar{X}, \Lambda)$. In particular, bounded rectangles in $\mathbb{R}^n$ will not be in this σ-algebra and therefore cannot be assigned conditional probability whereas they can be assigned probability by the full likelihood.

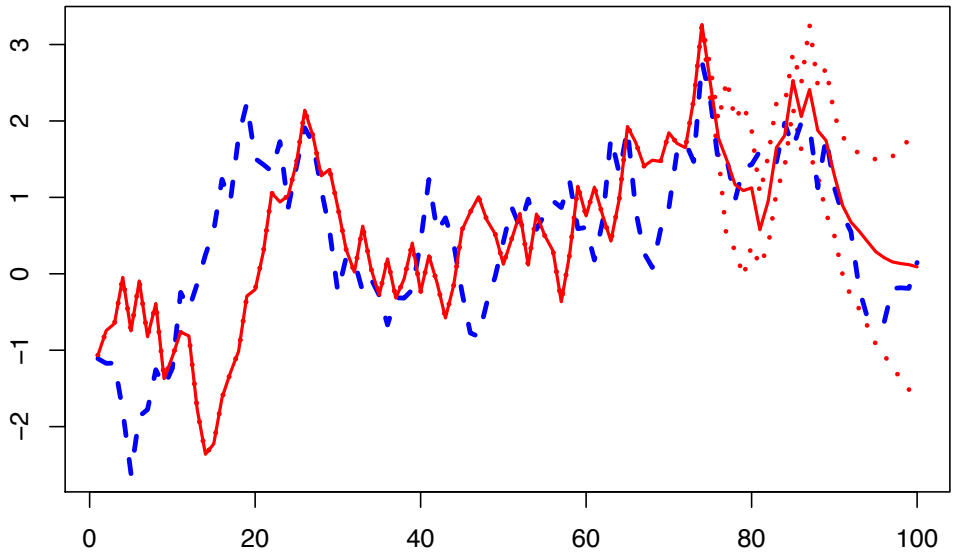

Figure 2. PF for Λ with Jacobian, Time Series Plot, SV Model. Data of length $T = 100$ was generated from a simulation of the model of Subsection 6.1 and $N = 5000$ particles computed using the algorithm described in Section 4.1 with a Jacobian term. The dashed blue line plots the simulated Λ for the last 50 time points. 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 (23) through (28); a one lag HAC estimator was used for (2).

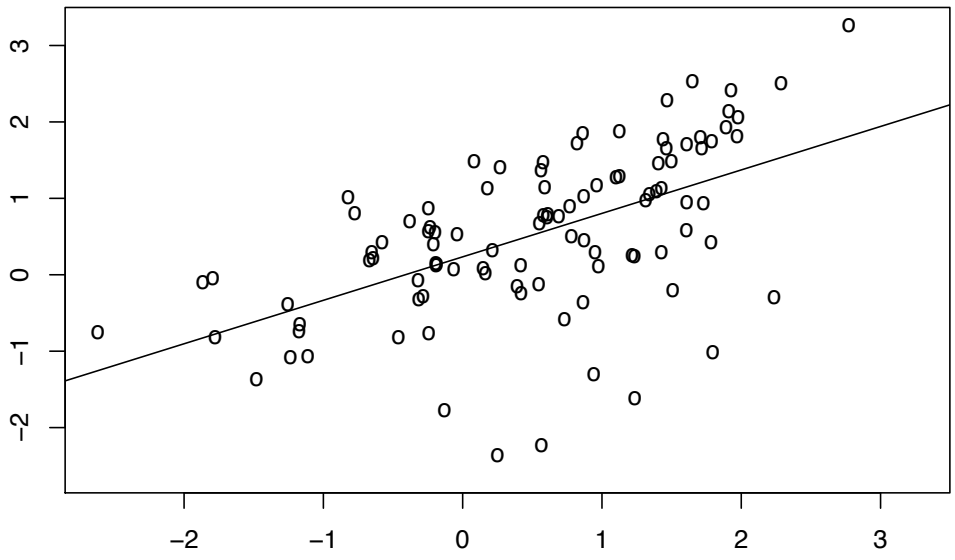

Figure 3. PF for Λ with Jacobian, Scatter Plot, SV Model. As for Figure 2 except that plotted is the mean of the particles vs. the simulated Λ.

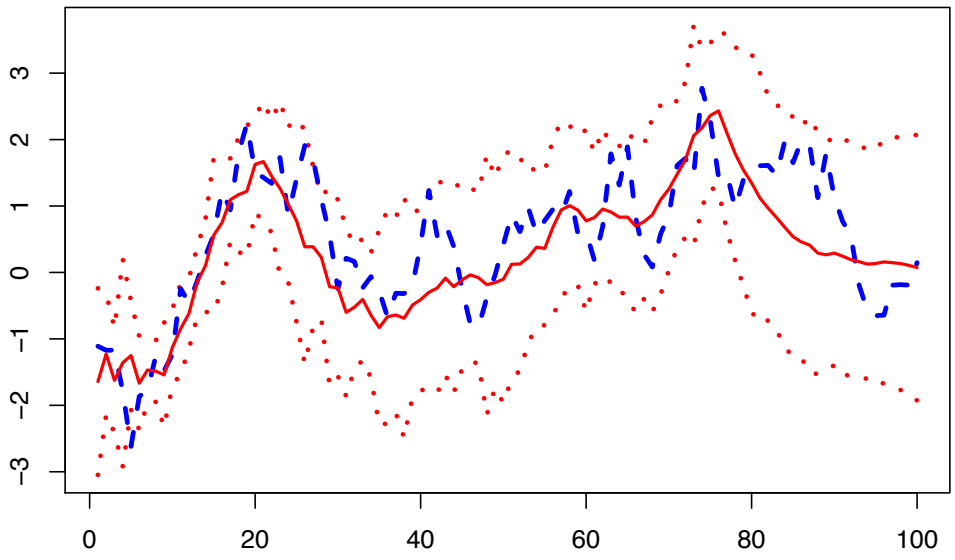

Figure 4. PF for Λ, without Jacobian, Time Series Plot. As for Figure 2 except that estimation is without a Jacobian term.

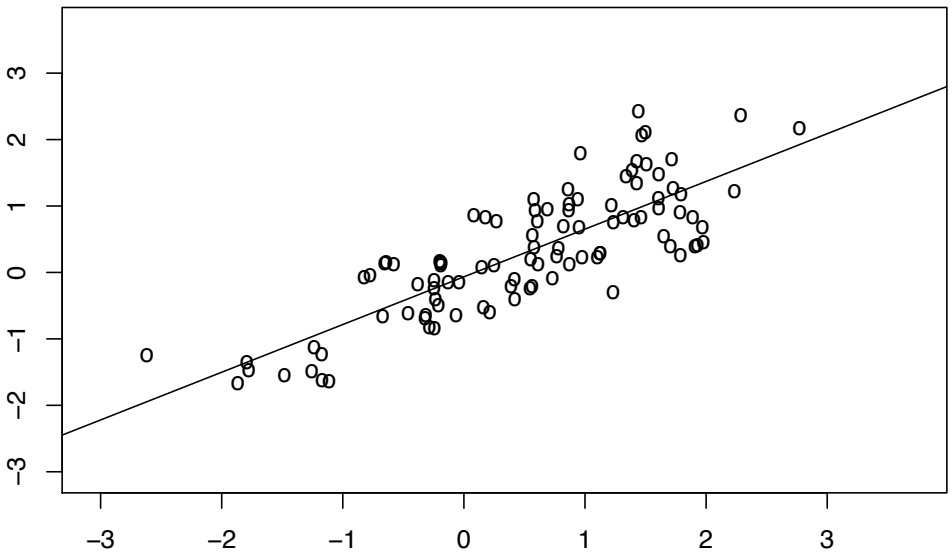

Figure 5. PF for Λ, without Jacobian, Scatter Plot, SV Model. As for Figure 4 except that plotted is the mean of the particles vs. the simulated Λ.

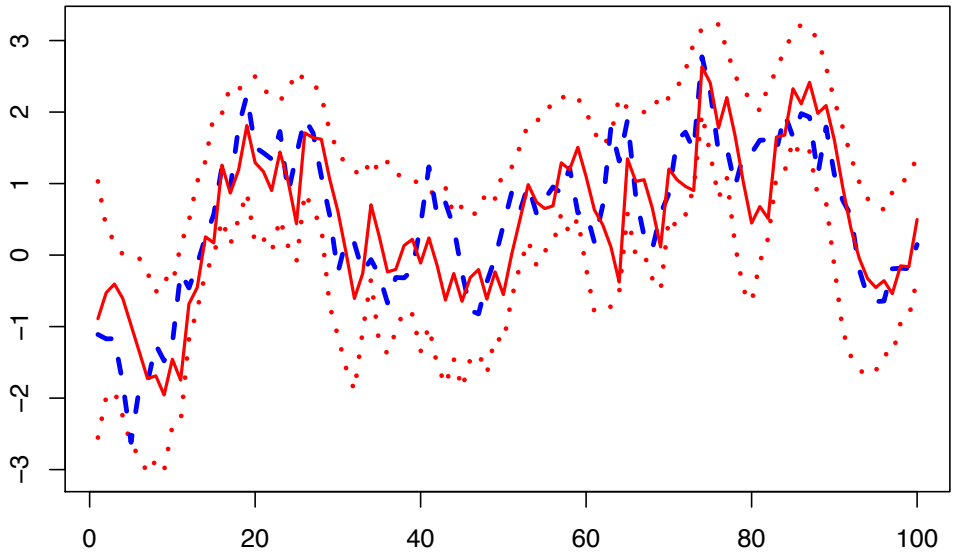

Figure 6. PE for Λ, Flurry-Shephard Method, Time Series Plot, SV Model. As for Figure 2 except that plotted is a filter, not a smooth, and weighting is by the measurement density, not GMM.

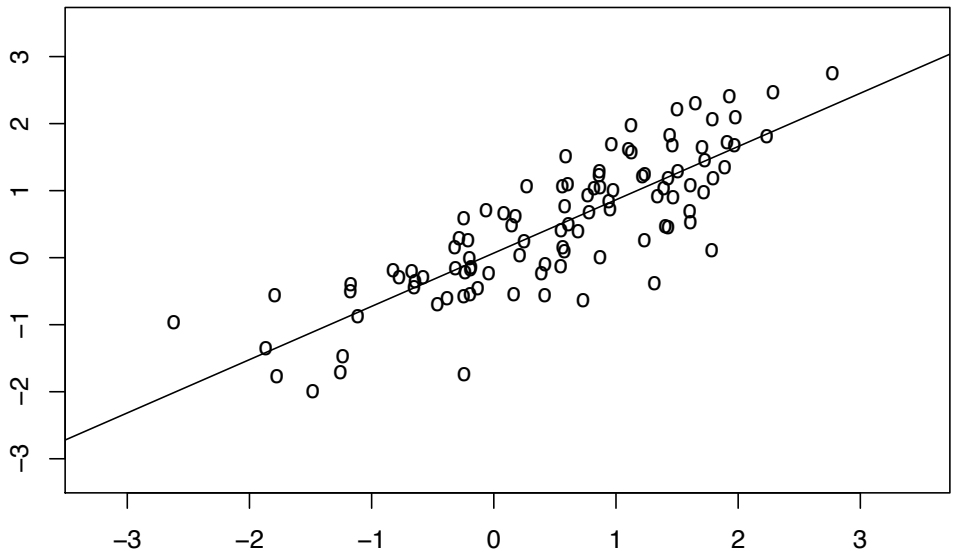

Figure 7. PF for Λ, Flurry-Shephard Method, Scatter Plot. As for Figure 6 except that plotted is the mean of the particles vs. the simulated Λ.

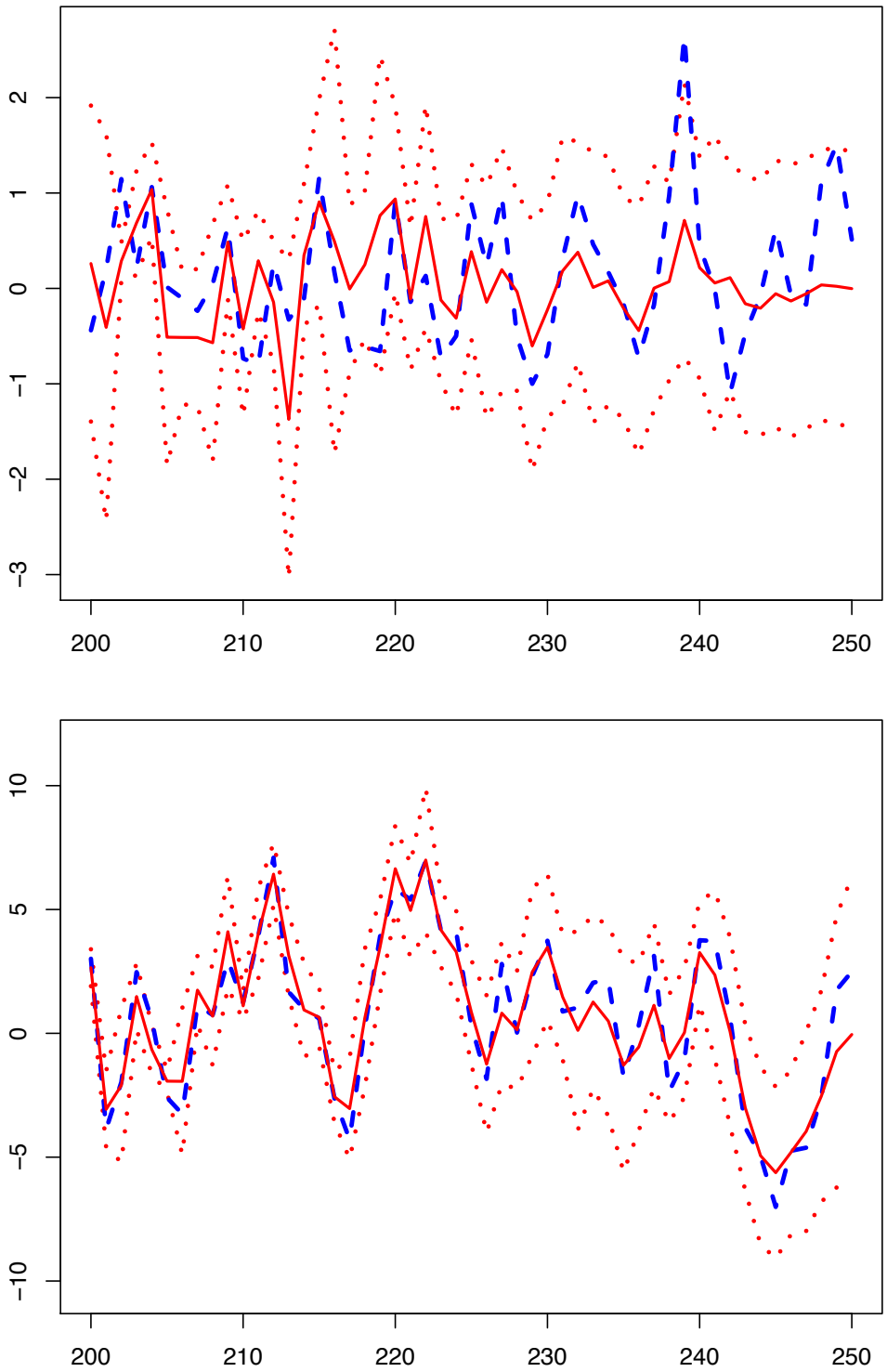

Figure 8. PF for Λ with Jacobian, Time Series Plot, DSGE Model. Data of length $T = 250$ was generated by simulating the model of Subsection 6.2 and $N = 10000$ particles were computed using the algorithm described in Section 4.1 with a Jacobian term. The dashed blue line in the upper panel plots the 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 (41) through (48); a two lag HAC estimator was used for (2).

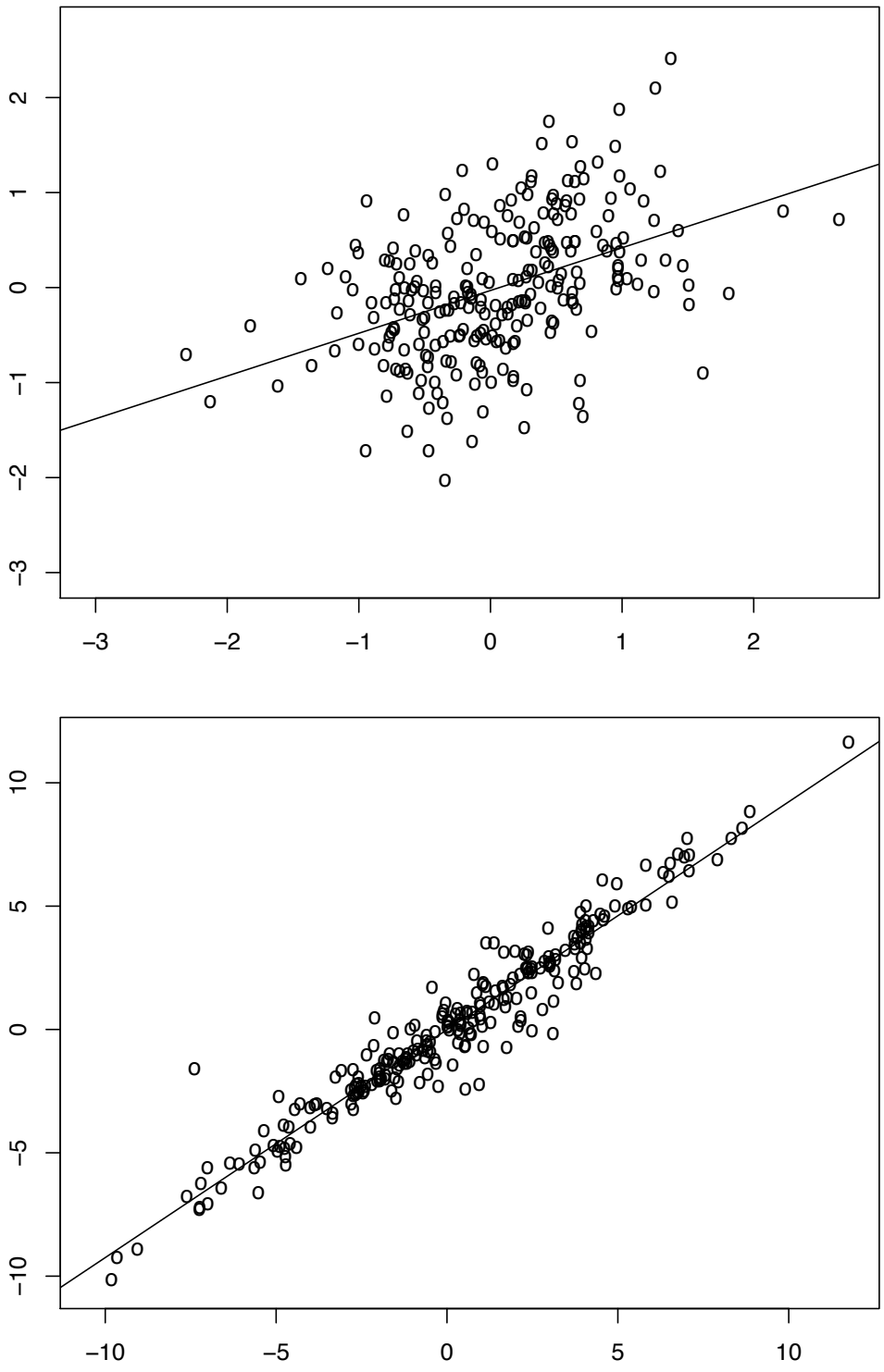

Figure 9. PF for Λ with Jacobian, Scatter Plot, DSGE Model. As for Figure 8 except that plotted is the mean of the particles vs. the simulated Λ for all 250 time points.

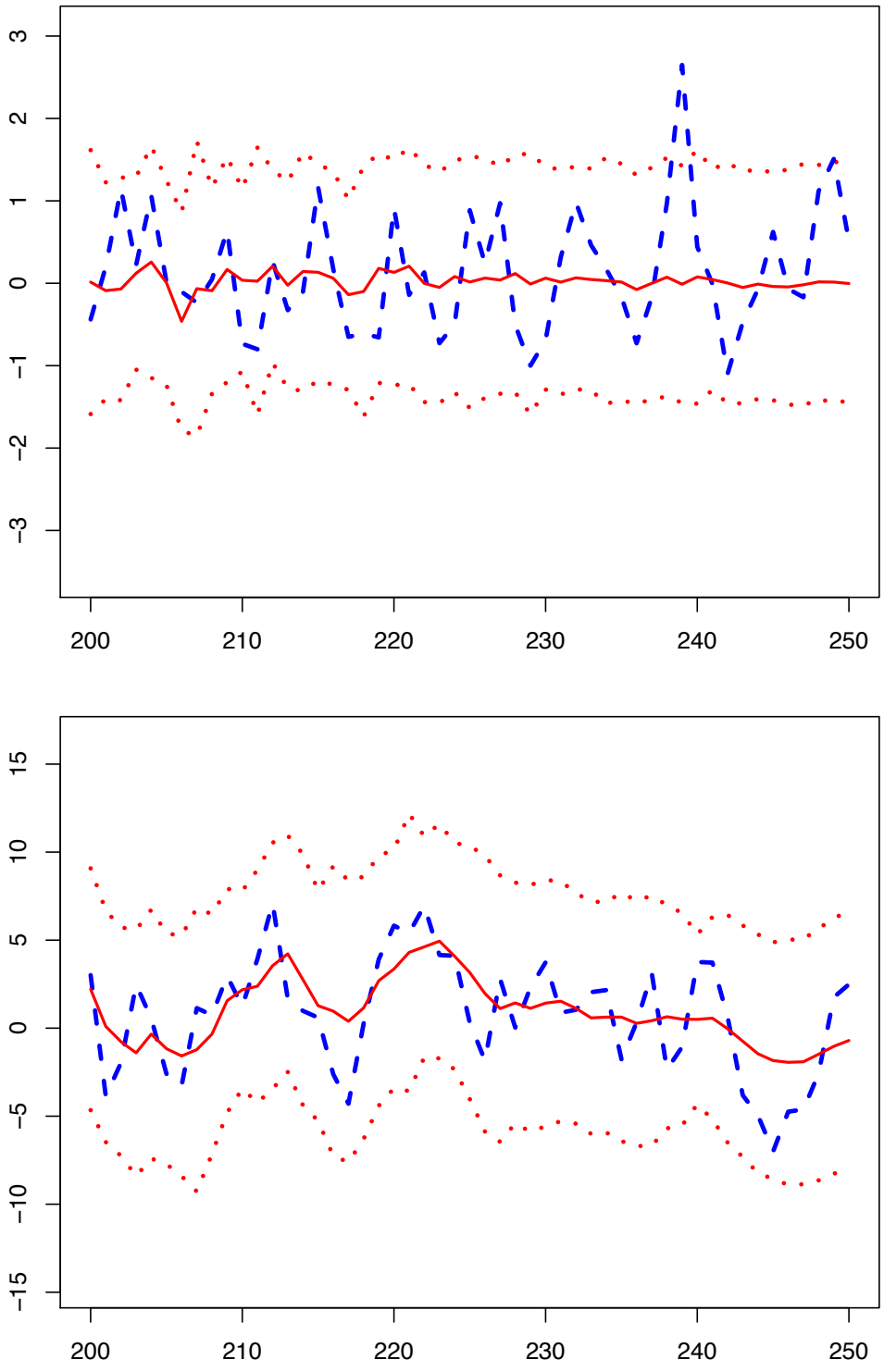

Figure 10. PF for Λ without Jacobian, Time Series Plot, DSGE Model. As for Figure 8 but without a Jacobian term.

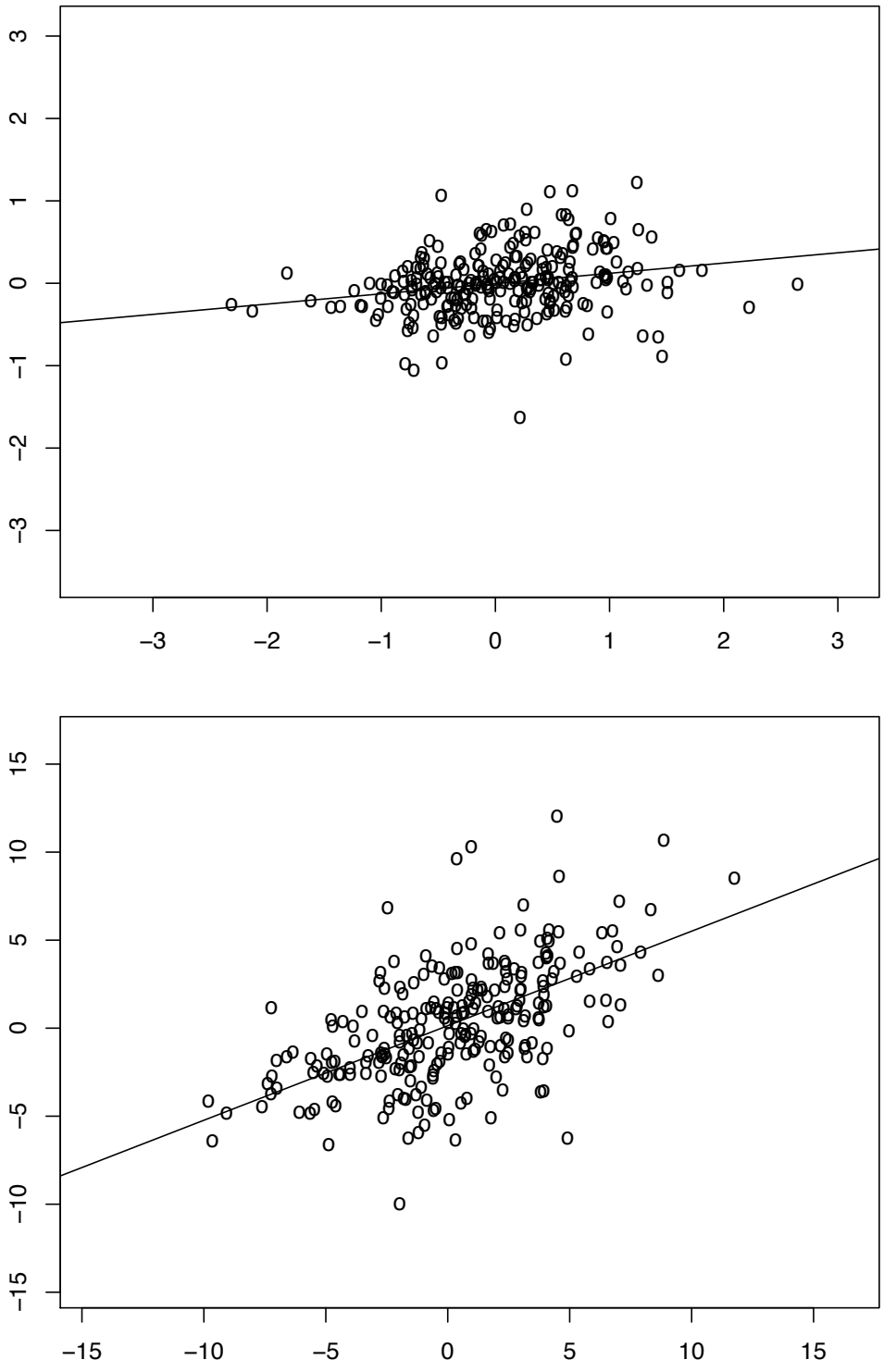

Figure 11. PF for Λ without Jacobian, Scatter Plot, DSGE Model. As for Figure 9 but without a Jacobian term.