Comments on GMM with Latent Variables[*]

A. Ronald Gallant
Duke University

Raffaella Giacomini
University College London

Giuseppe Ragusa
Luiss University

First draft: December 21, 2012
This draft: March 29, 2013

[*] Address correspondence to A. Ronald Gallant, Duke University, Fuqua School of Business, DUMC 90120, Durham NC 27708-0120, USA; phone 919-660-7700; email aronldg@gmail.com.

Abstract

We consider classical and Bayesian estimation procedures implemented by means of a set of conditional moment conditions that depend on latent variables. The latent variables evolve according to a Markovian transition density. Two main classes of applications are: 1) GMM estimation with time-varying parameters; and 2) estimation of nonlinear Dynamic Stochastic General Equilibrium (DSGE) models. The key idea is to base inference on an approximate likelihood that depends on conditional moment conditions. Bayesian estimation using this approach has received previous attention. The Bayesian results, which exploit some differences between Bayesian and frequentist inference, are summarized. Two methods for extending the Bayesian results to frequentist inference are discussed: 1) a particle filter approach. and, 2) a nonparametric sieve approach. At the present state of development, the former holds the most promise.

Keywords and Phrases: Generalized Method of Moments, Limited Information Maximum Likelihood, Latent Variables, Macro Models

JEL Classification: C32, C36, E27

1 Introduction

Limited information state space models – which we define as a set of moment conditions that depend on latent variables evolving according to a Markovian transition density – arise in several applications in economics. Examples of latent variables are time-varying parameters and structural shocks, which are often used in both partial and general equilibrium models to capture time variation and exogenous sources of persistence. The presence of latent variables poses some econometric challenges, which are usually addressed using a set of procedures that go under the name of state-space methods.

With state space methods, either Bayesian or frequentist, the standard method is to obtain the conditional density $p(X \mid \Lambda, \theta)$, where $X = (X_1, ..., X_T)$ and $\Lambda = (\Lambda_1, ..., \Lambda_T)$ denote the histories of observable and latent variables, respectively, and θ is the parameter of interest. Then one integrates to obtain the likelihood for observables $p(X|\theta) = \int p(X|\Lambda, \theta) p(\Lambda|\theta)\, d\Lambda$ using either analytic integration, numerical integration, or (particle) filtering (Fernandez-Villaverde and Rubio-Ramirez, 2006). This assumes knowledge of both densities inside the integral. If convenient analytic expressions are not available, these methods, as commonly implemented, involve numerical and analytic approximations that can be so inaccurate that econometric inference is misleading.

Without a likelihood, but for models that can be simulated, Bayesian methods can be implemented (Gallant and McCulloch, 2009), as can methods as efficient as maximum likelihood (Gallant and Tauchen, 1996), as well as GMM methods (Duffy and Singleton, 1993). However, the necessity of solving the model can introduce the same numerical compromises, such as log-linearization, that have invited criticism of likelihood based methods. Calibration methods suffer from this criticism for the same reason.

From the Bayesian perspective these criticisms are addressed by Gallant and Hong (2007). They form a limited information likelihood from a (continuously updated) GMM criterion. The method does not require an analytic expression for the density $p(\Lambda|\theta)$ but does require a prior opinion as to what it might be. In Gallant and Hong this opinion is expressed via a sieve representation of $p(\Lambda|\theta)$ and a prior on the coefficients of the sieve. Gallant and Hong heavily exploit some differences between Bayesian and frequentist inference with the consequence that their approach does not conveniently admit of an asymptotic analysis that could extend it from Bayesian to frequentist inference.

The contribution here is to modify the approach of Gallant and Hong so that it does admit of an asymptotic analysis. We review the Gallant and Hong approach in Section 3. In Section 4, we use the likelihood derived in Section 3 and directly eliminate the latent variables from the moment conditions by integration using a particle filter and then proceed along more traditional lines using the Chernozhukov and Hong (2003) method to optimize the likelihood. In Section 5, using an alternative extension of Gallant and Hong, we nonparametrically estimate the realized history of the latent process, which has the side effect of eliminating the need for a prior opinion regarding the probability law of the latent process. Inference is based on a GMM criterion, which is also nonparametric in the sense that no distributional assumptions are required beyond existence of moments.

2 Examples

2.1 A Stochastic Volatility Model

A simple example of the foregoing is a stochastic volatility 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

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

The moment conditions used with this model are:

$$
\begin{aligned}
h_1 &= (X_t - \rho X_{t-1})^2 - [\exp(\Lambda_t)]^2 && (1)\\
h_2 &= |X_t - \rho X_{t-1}||X_{t-1} - \rho X_{t-2}| - \left(\frac{2}{\pi}\right)^2 \exp(\Lambda_t)\exp(\Lambda_{t-1}) && (2)\\
&\vdots \\
h_{L+1} &= |X_t - \rho X_{t-1}||X_{t-L} - \rho X_{t-L-1}| - \left(\frac{2}{\pi}\right)^2 \exp(\Lambda_t)\exp(\Lambda_{t-L}) && (3)\\
h_{L+2} &= X_{t-1}(X_t - \rho X_{t-1}) && (4)\\
h_{L+3} &= \Lambda_{t-1}(\Lambda_t - \phi\Lambda_{t-1}) && (5)\\
h_{L+4} &= (\Lambda_t - \phi\Lambda_{t-1})^2 - \sigma^2 && (6)
\end{aligned}
$$

2.2 A Dynamic Stochastic General Equilibrium Model

This example is taken from Del Negro and Schorfheide (2008). We need to have a model with an exact analytical solution to generate reliable 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. A solid argument in the working paper that the model is identified is what distinguishes the version we use from the others.

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 distributed as

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

The first order conditions are

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

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

The solution for the endogenous variables is

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

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

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

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

The moment conditions that we shall use are

$$
\begin{aligned}
h_1 &= (z_{t+1} - \rho_z z_t)^2 - \sigma_z^2 && (7)\\
h_2 &= z_t(z_{t+1} - \rho_z z_t) && (8)\\
h_3 &= (\phi_{t+1} - \rho_\phi \phi_t)^2 - \sigma_\phi^2 && (9)\\
h_4 &= \phi_t(\phi_{t+1} - \rho_\phi \phi_t) && (10)\\
h_5 &= (w_{t+1} - \rho_\lambda w_t)^2 - \sigma_\lambda^2 && (11)\\
h_6 &= w_t(w_{t+1} - \rho_\lambda w_t) && (12)\\
h_7 &= w_{t+1} - (1+\nu)y_{t+1} - \phi_{t+1} && (13)\\
h_8 &= y_{t+1} + \pi_{t+1} + z_{t+1} - y_t - \frac{1}{\beta}\pi_t && (14)
\end{aligned}
$$

Some additional moment conditions one might consider are

$$
\begin{aligned}
h_9 &= w_t h_8 && (15)\\
h_{10} &= y_t h_8 && (16)\\
h_{11} &= \pi_t h_8 && (17)\\
h_{12} &= z_t h_8 && (18)\\
h_{13} &= \phi_t h_8 && (19)
\end{aligned}
$$

3 Bayesian Methods

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). They used the method to extract the monthly and annual pricing kernels from a panel of equity and fixed income securities using a GMM criterion function derived from Euler conditions. We will introduce their ideas with a simple example, follow that with the general case, and then illustrative with their application.

3.1 A Simple 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 convenience, is the distribution that we shall use for Z. Even without assuming that the X_i are normal an assumption that Z is normal is often reasonable. Thus, $\bar{X}$ and Λ have joint density

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

The mathematical justification for this assertion is in Gallant and Hong. Joint probability on $(\bar{Y}, \Lambda)$ can only be assigned to (the smallest σ-algebra containing all) sets bounded by 45 degree lines. An example is the set labeled $A_{(\bar{Y},\Lambda)}$ in Figure 1. The conditional probability for a set such as that labeled $C_{(\bar{Y}|\Lambda)}$ 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; details are in Gallant and Hong. 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 $p(\bar{X}, \Lambda)$. These sets 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.

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.

3.2 Bayesian Estimation

We now consider the general case. As above, let $X = (X_1, ..., X_T)$ and $\Lambda = (\Lambda_1, ..., \Lambda_T)$ denote histories of observable and latent variables, respectively, and θ be the parameter of interest. Let $h : \mathbb{R}^{d_x} \times \mathbb{R}^{d_\lambda} \times \mathbb{R}^{d_\beta} \to \mathbb{R}^M$ be a set of M moment conditions involving d_x observable variables X_t and d_λ latent variables Λ_t. A limited information state space model is defined as

$$E\left[h(X_{t+1}, \Lambda_{t+1}, \beta) \,|\, \mathcal{I}_t\right] = 0 \tag{20}$$

$$\Lambda_{t+1} \sim P(\Lambda_{t+1} \,|\, \Lambda_t, \gamma) \tag{21}$$

where $\mathcal{I}_t = \{\ldots, X_1, \Lambda_1, X_2, \Lambda_2, \ldots, X_t, \Lambda_t\}$ is the information set at time t. Let $\theta = (\beta, \gamma)$ with any redundancies eliminated and let θ_0 denote the true value of θ. The system in (20) and (21) implies a set of M unconditional moment conditions

$$E[g(X_{t+1}, \Lambda_{t+1}, \theta_0)] = 0. \tag{22}$$

As above, the likelihood $p(X, \Lambda, \theta)$ is based on a transformation of the GMM objective function:

$$p(X, \Lambda, \theta) = (2\pi)^{-M/2} \exp\left\{ -\frac{1}{2} g_T(X, \Lambda, \theta)' \left[\Sigma(X, \Lambda, \theta)\right]^{-1} g_T(X, \Lambda, \theta) \right\}. \tag{23}$$

where

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

$$\begin{aligned} \Sigma(X, \Lambda, \theta) &= \frac{1}{T} \sum_{t=1}^{T} \left[\widetilde{g}(X_t, \Lambda_t, \theta)\right] \left[\widetilde{g}(X_t, \Lambda_t, \theta)\right]', \\ \widetilde{g}(X_t, \Lambda_t, \theta) &= g(X_t, \Lambda_t, \theta) - \frac{1}{\sqrt{T}} g_T(X, \Lambda, \theta). \end{aligned} \tag{24}$$

Depending on the nature of the dynamics and moments, a HAC estimator (Gallant, 1987, p. 445) of $\Sigma(X, \Lambda, \theta)$ may have to be substituted for (24).

The likelihood $p(X, \Lambda, \theta)$ is used in the usual way for Bayesian inference. That is, one specifies a prior $p(\Lambda, \theta)$ and draws from the posterior using Markov Chain Monte Carlo (MCMC) (Gamerman and Lopes, 2006). As the dimension of Λ can be large, this can be computationally intensive. Perhaps the best known example of this computational approach in the econometrics literature is Jacquier, Polson, and Rossi (1994). Because Bayesian inference is subjective, one does not have to regard $p(X, \Lambda, \theta)$ as an approximation to the likelihood justified by asymptotics, although most would.

3.3 Illustrative Example

Any asset pricing kernel $\{\Lambda_t\}$ will satisfy these Euler equations

$$1 = \mathcal{E}_t \left(\Lambda_{t+1} X_{s,t+1} \right), \tag{25}$$

where $X_{s,t}$ is the gross return on asset s at time t. Gallant and Hong used 551 monthly returns on the Fama and French (1993) portfolios and U.S. Treasury debt of ten year, one year, and thirty day maturities as data to implement the moment conditions and these data lagged as well as aggregate stock returns, consumption growth, and labor income growth as instruments. The Euler equations interacted with the instruments comprise the moment conditions $g_T(X, \Lambda, \theta)$ above. There are enough moment conditions to overidentify $\{\Lambda_t\}$, although overidentification is not required as seen from Jacquier, Polson, and Rossi (1994). The variance matrix $\Sigma(X, \Lambda, \theta)$ has a factor structure with known eigenvectors, which dramatically reduces the dimensionality of the computations; θ contains the eigenvalues of $\Sigma(X, \Lambda, \theta)$. These moment conditions define the likelihood $p(X, \Lambda, \theta)$ given by (23). The prior used by Gallant and Hong has the form

$$\left[\prod_{t=1}^{n} f(\Lambda_{t+1} | \Lambda_t, \ldots, \Lambda_1, \eta) \right] f(\Lambda_1 | \eta)\, p(\eta)$$

where $f(\Lambda_{t+1} | \Lambda_t, \ldots, \Lambda_1, \eta)$ is the sieve for the law of motion of Λ of Gallant and Nychka (1987) and $p(\eta)$ mildly tilts the law of motion toward the Bansal and Yaron (2004) calibration of a long-run risks economy. Draws from the posterior are obtained by MCMC using the code at http://public.econ.duke.edu/webfiles/arg/emm. The mean and standard deviation of the posterior for Λ are shown in Figure 2.

(Figure 2 about here)

4 Particle Filter Methods

Recall from Subsection 3.2 that we have two jointly distributed histories under consideration, the observed

$$X = (X_1, ..., X_T)$$

and the unobserved

$$\Lambda = (\Lambda_1, ..., \Lambda_T).$$

We shall denote partial histories by

$$X_{1:t} = (X_1, ..., X_t)$$

and

$$\Lambda_{1:t} = (\Lambda_1, ..., \Lambda_t),$$

respectively. The joint distribution depends on an unknown parameter θ. A particle filter is a computationally efficient method for drawing from the conditional distribution of Λ given X with θ specified. Therefore, given a set of moment conditions

$$g_T(X, \Lambda, \theta)$$

that depend on the joint history, the draws can be used to compute

$$g_T(X, \theta) = \mathcal{E}[g_T(X, \Lambda, \theta) \,|\, X].$$

By the law of iterated expectations, $\mathcal{E} g_T(X, \theta) = 0$. Therefore, $g_T(X, \theta)$ can be used for GMM estimation in the usual way. In this section we shall describe the computational procedure and apply it to the examples in Section 2. The theoretical justification follows in Subsection 4.4.

The setup is as in Subsection 3.2 with one important difference. The density (23) is the density of

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

as dictated by the theory developed in Gallant and Hong (2007). In frequentist inference one has some flexibility as to what density to use and one might consider density of $g_T(X,\Lambda,\theta)$, which is

$$\widehat{p}(X,\Lambda,\theta) = [2\pi\ \det\Sigma(X,\Lambda,\theta)]^{-M/2}\exp\left\{-\frac{1}{2}g_T(X,\Lambda,\theta)'\,[\Sigma(X,\Lambda,\theta)]^{-1}\,g_T(X,\Lambda,\theta)\right\}. \qquad (26)$$

The reason is that (23) is a continuously updated GMM estimator, which has well known tail problems in frequentist inference. The determinant term in (26) may alleviate the tail problem. We consider this issue in the numerical experiments that we conduct. When computing the above quantities from partial histories we shall substitute the subscript t for T.

For given θ, the particle filter algorithm is as follows.

1. Initialization.
   - (a) Set T_0 to the minimum sample size required to compute $g_t(X_{1:t},\Lambda_{1:t},\theta)$.
   - (b) For $i=1,\ldots,N$ sample $(\Lambda_1^{(i)},\Lambda_2^{(i)},\ldots,\Lambda_{T_0}^{(i)})$ from $p(\Lambda_t|\Lambda_{t-1},\gamma)$.
   - (c) Set t to T_0+1.
   - (d) Set $\Lambda_{1:t-1}^{(i)} = (\Lambda_1^{(i)},\Lambda_2^{(i)},\ldots,\Lambda_{T_0}^{(i)})$
2. Importance sampling step.
   - (a) For $i=1,\ldots,N$ sample $\tilde{\Lambda}_t^{(i)}$ from $p(\Lambda_t|\Lambda_{t-1}^{(i)})$ and set $\tilde{\Lambda}_{1:t}^{(i)} = (\Lambda_{0:t-1}^{(i)},\tilde{\Lambda}_t^{(i)})$.
   - (b) For $i=1,\ldots,N$ compute weights $\tilde{w}_t^{(i)} = \hat{p}(X_{1:t},\tilde{\Lambda}_{1:t}^{(i)},\theta)$.
   - (c) Normalize the weights, i.e., $\tilde{w}_t^{(i)} \leftarrow \tilde{w}_t^{(i)}/\sum_{t=1}^{N}\tilde{w}_t^{(i)}$.
3. Selection step.
   - (a) For $i=1,\ldots,N$ sample with replacement from the set $\{\tilde{\Lambda}_{1:t}^{(i)}\}$ according to the normalized weights $\tilde{w}_t^{(i)}$ to obtain a new set of particles $\{\Lambda_{1:t}^{(i)}\}$ that have equal weight.
   - (b) If $t<T$, increment t and go to Step 2; else go to Step 4.
4. Finalize.

(a) Compute

$$\bar{g}_T(\theta) = \frac{1}{N} \sum_{i=1}^{N} g_T(X, \Lambda_{1:T}^{(i)}, \theta)$$

(b) Compute

$$\bar{\Sigma}_T(\theta) = \frac{1}{N} \sum_{i=1}^{N} \Sigma_T(X, \Lambda_{1:T}^{(i)}, \theta)$$

(c) Set

$$\bar{p}_T(\theta) = (2\pi \det \bar{\Sigma}_T(\theta))^{-M/2} \exp \left\{ -\frac{1}{2} \bar{g}_T'(\theta) \left[\bar{\Sigma}_T(\theta) \right]^{-1} \bar{g}_T(\theta) \right\}$$

The density $\bar{p}_T(\theta)$ can be treated as a likelihood and standard maximum likelihood formulae can be used for statistical inference. However, there is the complication that even if the particle filter commences with the same seed at Step 1 for each θ, $\bar{p}_T(\theta)$ will not be differentiable with respect to θ due to Step 3. This is a well known problem and there are various ways to deal with it (Pitt, 2002) and (Flury and Shephard, 2010). The first involves smoothing the selection method; it is difficult to implement when there is more than one latent variable. The second uses a variable seed instead of a fixed seed to start the particle filter algorithm every time θ is changed by an MCMC method with N increased as necessary to control the rejection rate of the chain. An example of its use appears in Table 1. Neither method gets around the main problem with particle filter methods which is that they are computationally intensive. The Flury and Shephard method can make the problem worse or better; it's a matter of trial and error.

We propose an EM method (Dempster, Laird, Rubin, 1977) that dramatically reduces computational cost and that we believe is new to this paper. Our solution is to use the particle filter algorithm as the expectation part of an EM method. We use the Chernozhukov-Hong (2003) method as implemented at http://public.econ.duke.edu/webfiles/arg/emm but only at every 50th MCMC draw do we run the particle filter algorithm. At all the others we reuse the particles from the previous computation. Runtimes decrease by about 1/50 because the particle filter is nearly the entire computational cost. (The number 50 is larger than 5 times the number of parameters for our examples so that all can be expected to move between E steps under a move-one-at-time random walk proposal.) Also, there is less need for a large number of particles because there is no essential need for smoothness with our proposed approach. We refer to it as the EM PF GMM method hereafter.

With this EM approach it is necessary that the moment equations would be sufficient to identify all the parameters of the model were both X and Λ observed because this is essentially the situation during the M step of the EM algorithm. For instance, for the stochastic volatility model of Subsection 4.2 the moments (5) and (6) are necessary whereas they would not be if one worked by analogy with conventional particle filter methods with an analytic measurement density available. A conventional particle filter with analytic measurement density is effectively only using moment conditions (1) and (4). Moment conditions (2) through (3) are overidentifying conditions.

Strictly speaking, what we propose is not an EM algorithm because neither the E step nor the M step is guaranteed to increase the objective function. Present experience suggests that this doesn't matter because an MCMC chain can tolerate some choppiness. What MCMC cannot tolerate is valleys that it cannot cross, which is what seems to happen when the E step is eliminated (by running the particle filter at every draw of the MCMC chain).

It is helpful to regularize inversion of the weighting matrix (48). If the condition number of the weighting matrix (ratio of smallest singular value to the largest) falls below a preset value η (e.g. $\eta = 10^{-13}$) 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. If the inference strategy were Bayesian, the idea of Andrieu, Douced, and Holenstein (2010) of using the particle filter as a proposal within an MCMC chain, which requires actually computing the transition density induced by the particle filter, would be of interest but it does not seem to be compatible with the view of this section that Λ is a variable to be eliminated by integration rather than a parameter to be estimated.

4.1 Caveat

The computations in Subsections 4.2 and 4.3 that follow are suggestive of eventual success but are defective in that they contain undissipated transients and we have not yet gotten the chains to mix adequately. In some instances the chains have drifted to unreasonable parameter values.

4.2 Stochastic Volatility Example

Estimates of θ for the stochastic volatility example of Subsection 4.2 are shown in Table 1 for three methods: EM PF GMM with a Jacobian term, without a Jacobian term, and using the Flury and Shephard (2010) estimator.

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 3 and as a scatter plot in Figure 4 for the case when a Jacobian term is included and as Figures 5 and 6 when it is not. The plots for the Flury and Shephard estimator are Figures 7 and 8. In the particle filter vernacular, the EM PF GMM estimator is computed from a smooth whereas the Flury and Shephard estimator is computed from a filter; accordingly, the plots shown for the EM PF GMM estimator are smooths whereas the plots shown of the Flury-Shephard estimator are filters.

The Flury and Shephard estimator is not strictly comparable because it is a Bayesian estimator and requires a likelihood, which, in turn, requires the use of numerical methods whose quality can be dubious in DSGE models, as discussed in Section 1. In the stochastic volatility example an analytic likelihood is available. We find that the essential idea of Flury and Shephard of letting the seed be random does not work for our GMM objective function because the number of particles has to be so large to control the rejection rate that computational cost becomes a serious issue. Moreover, the proof strategy that justifies it does not apply to our estimator.

(Table 1 about here)

(Figure 3 about here)

(Figure 4 about here)

(Figure 5 about here)

(Figure 6 about here)

(Figure 7 about here)

(Figure 8 about here)

4.3 Dynamic Stochastic General Equilibrium Example

Applying EM PF GMM method both with and without a Jacobian term to the DSGE model of Subsection 2.2, we obtain the estimates of θ shown in Table 2.

Applying the particle filter at the true of θ and $N = 10000$, we obtain the estimate of Λ shown as a time series plots in Figures 9 and 11 and as a scatter plots in Figures 10 and 12.

(Table 2 about here)

(Figure 9 about here)

(Figure 10 about here)

(Figure 11 about here)

(Figure 12 about here)

4.4 Theoretical Justification

For simplicity, we shall drop the Jacobian term and use the likelihood discussed in Subsections 3.1 and 3.2. The case with the Jacobian term is analogous.

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 theta 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 assume, as in the example of Subsection 3.1, 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 $X_{1:t} \rightarrow X^*_{1:t} \rightarrow z_t$ in the opposite direction. Note that if $X^*_{1:t} \rightarrow z^*_t$ then $z^*_t \rightarrow X^*_{1:t}$; therefore, for convenience, we will always choose $X^o_{1:t}$ as the $X^*_{1:t}$ for its image so that $X^o_{1:t} \rightarrow z^o_t \rightarrow X^o_{1:t}$.

The following three 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

$$g(z_{1:t}) = f[X_{1:1}(z_1,\Lambda_{1:t},\theta), X_{1:2}(z_2,\Lambda_{1:t},\theta),\ldots,X_{1:t}(z_t,\Lambda_{1:t},\theta)] \tag{27}$$

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

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

(3) The function $\bar{p}_T(\theta)$ returned at the end of Step 4 of the particle filter has the form of Equation 27.

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

$$p(z_t\,|\,\Lambda_{1:t},\theta) = n\left\{[Z_t[X_{1:t}(z_t,\Lambda_{1:t},\theta),\Lambda_{1:t},\theta]\,|\,0,I\right\} \tag{28}$$

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 acutal distribution of $Z_T(X^o_{1:T}, \Lambda, \theta)$ and $\psi(\cdot)$ its density function. We assume that $\psi_T(\cdot)$ converges in distribution to the standard normal distribution $\phi(\cdot)$, with density $\phi(\cdot)$, for large T. Let

$$\begin{aligned} u_T^{(i)} &= \phi(z_T^{(i)})\, p(\Lambda \,|\, \theta) & (29)\\ U_T &= \int \phi(Z_T(X^o_{1:T}, \Lambda, \theta))\, p(\Lambda \,|\, \theta)\, d\Lambda & (30)\\ v_T^{(i)} &= \psi_T(z_T^{(i)})\, p(\Lambda \,|\, \theta) & (31)\\ V_T &= \int \psi_T(Z_T(X^o_{1:T}, \Lambda, \theta))\, p(\Lambda \,|\, \theta)\, d\Lambda & (32) \end{aligned}$$

where

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

Using (29) through (32) to construct importance sampling weights, we have

$$\frac{1}{N} \sum_{i=1}^{N} \frac{v_T^{(i)}}{u_T^{(i)}} \frac{U_T}{V_T}\, g_T(X^o e :_{1:T}, \Lambda_{1:T}^{(i)}, \theta) = \frac{U_T}{V_T} \frac{1}{N} \sum_{i=1}^{N} \frac{\psi_T(z_T^{(i)})}{\phi(z_T^{(i)})}\, g_T(X^o_{1:T}, \Lambda_{1:T}^{(i)}, \theta) \qquad (33)$$

is an approximation to

$$\int g_T(X^o_{1:T}, \Lambda, \theta)\, f_T(\Lambda \,|\, z_{1:T}, \theta)\, d\Lambda \qquad (34)$$

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

We shall first show that

$$\frac{U_T}{V_T} \frac{1}{N} \sum_{i=1}^{N} g_T(X^o_{1:T}, \Lambda_{1:T}^{(i)}, \theta) \qquad (35)$$

also approximates (34) for large N and T.

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

$$\frac{U_T}{V_T} \int I\{Z_T(X^o_{1:T}, \Lambda, \theta) \in (a_0, b_0]\}\, g_T(X^o_{1:T}, \Lambda, \theta)\, f_T(\Lambda \,|\, z_{1:T}, \theta)\, d\Lambda \qquad (36)$$

approximates (34) 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

$$\frac{U_T}{V_T} \frac{1}{N} \sum_{i=1}^{N} I\{Z_T(X^o_{1:T}, \Lambda, \theta) \in (a_0, b_0]\}\, \frac{\psi_T(z_T^{(i)})}{\phi(z_T^{(i)})}\, g_T(X^o_{1:T}, \Lambda_{1:T}^{(i)}, \theta) \qquad (37)$$

approximates (36) 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 (35) approximates (34) 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

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

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

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

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

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

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

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

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

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

Regularity conditions sufficient for particles to be draws $\Lambda^{(i)}_{1:T}$ from the density $p(\Lambda_{1:T} \mid z_{1:T}, \theta)$ are in Andrieu, Douced, and Holensteing (2010) and the references therein. They are mild, requiring that the weights at Step 2a be bounded and that multinomial resampling be used, which is the scheme used at Step 3a. We now have pointwise convergence in θ which is adequate for our purposes. With some more work that we have yet to do, consistency and asymptotic normality follow from Gallant (1987, Chapter 7).

5 Nonparametric Methods

In this section we nonparametrically estimate the realized history of the latent process, which has the side effect of eliminating Gallant and Hong's (2007) need for a prior opinion regarding the probability law of the latent process. As inference is based on a GMM criterion, the approach is completely nonparametric in the sense that no distributional assumptions are required beyond existence of moments.

To motivate our approach, consider the standard paradigm regarding latent variables: One has an abstract probability space from which Nature draws that leads via a diffusion to a trajectory for the latent variables that is continuous but nondifferentiable. The economic agent samples this trajectory at discrete intervals, say monthly. The probability law of the sampled sequence can usually be given a convenient discrete time representation such as an autoregression. We show that under standard assumptions on the probability law of the latent process there corresponds to the sampled sequence a function that has an invertible Fourier transform. The Fourier transform of this function function is the object that we shall estimate nonparametrically. An estimate of the transform can be inverted to obtain an estimate of the history of the latent process that governed the agent's decisions.

5.1 The Fourier Transform of the Latent History

In this section we describe the Fourier transform of the history of the latent process that we shall use in the sequel. A reference is Rahman (2011).

Let $\{y_i\}_{i=1}^{\infty}$ be a real-valued random process defined on the positive integers. Let $Y(\cdot)$ mapping $\mathbb{R}$ into $\mathbb{R}$ be the left-continuous random process $Y(t) = \sum_{i=1}^{\infty} i^{-3} y_i I_{[i,i+1)}(t)$ where

$I_A(t)$ denotes the indicator function; i.e., $I_A = 1$ if $t \in A$ and 0 else.

LEMMA 1 If $\frac{1}{n}\sum_{i=1}^{n} |y_i|^p$ converges almost surely as n tends to infinity, then $Y \in L_q$ almost surely for all $1 \le q \le p$.

Proof Let $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} |y_i|^p$. Given $\epsilon > 0$ there is an N such that $n > N$ implies $m - \epsilon < \bar{y}_n < m + \epsilon$ for some $m \ge 0$ except on an event that occurs with probability zero. Because $i^{-3p}|y_i|^p = i^{-3p+1}\bar{y}_i - i^{-3p+1}\bar{y}_{i-1} + i^{-3p}\bar{y}_{i-1}$,

$$-2\epsilon \sum_{i=1}^{n} i^{-3p-1} + (m-\epsilon)\sum_{i=1}^{n} i^{-3p} \le \sum_{i=1}^{n} i^{-3p}\,|y_n|^p \le 2\epsilon \sum_{i=1}^{n} i^{-3p-1} + (m+\epsilon)\sum_{i=1}^{n} i^{-3p}.$$

Therefore $\int_{-\infty}^{\infty} |Y(t)|^p\,dt = \lim_{n\to\infty}\sum_{i=1}^{n} i^{-3p}\,|y_i|^p \approx m\int_{1/2}^{\infty} x^{-3p}\,dx$. On a probability space, almost sure convergence of $\frac{1}{n}\sum_{i=1}^{n}|y_i|^q$ for $q = p$ implies it for $1 \le q \le p$. □

For $p = 2$ Lemma 1 implies that Y is in both L_1 and L_2. Such functions have a Fourier transform

$$\hat{Y}(\omega) = \int_{-\infty}^{\infty} Y(t)e^{-i\omega t}\,dt$$

that is in L_2 and can be inverted to obtain y_t using

$$y_t = \frac{t^3}{2\pi}\int_{-\infty}^{\infty} \hat{Y}(\omega)e^{i\omega t}\,d\omega. \tag{38}$$

Write

$$\hat{Y}(\omega) = A(\omega) + iB(\omega).$$

Because y_t is real, we must have that $A(\omega)$ is an even function and $B(\omega)$ is an odd function.

The two functions A and B in L_2 will be the objects of interest in later sections. They need a sieve representation. One has considerable flexibility because there are many basis functions for L_2. Of those for which (38) has a known analytic expression, the basis functions $\omega^n e^{-\omega^2/4}$ are convenient for representing even and odd functions:

$$\begin{aligned} A(\omega) &= \sqrt{\pi}\sum_{k=0}^{\infty} a_k\,\omega^{2k}e^{-\omega^2/4} \qquad (39)\\ B(\omega) &= \sqrt{\pi}\sum_{k=0}^{\infty} b_k\,\omega^{2k+1}e^{-\omega^2/4}. \end{aligned}$$

The corresponding sieve is

$$\begin{aligned} A_K(\omega) &= \sqrt{\pi}\sum_{k=0}^{K} a_k\,\omega^{2k}e^{-\omega^2/4} \qquad (40) \\ B_K(\omega) &= \sqrt{\pi}\sum_{k=0}^{K} b_k\,\omega^{2k+1}e^{-\omega^2/4}. \end{aligned}$$

Application of (38) yields

$$y_t = t^3\,e^{-t^2}\sum_{k=0}^{K}(-1)^k\left[a_k\,H_{2k}(t) - b_k\,H_{2k+1}(t)\right], \qquad (41)$$

where $H_n(\cdot)$ denotes a Hermite polynomial of degree n. The Hermite polynomials can be computed from the recursion

$$\begin{aligned} H_0(x) &= 1 \\ H_1(x) &= 2x \\ H_n(x) &= 2xH_{n-1}(x) - 2(n-1)H_{n-2}(x). \end{aligned}$$

Lemma 1 remains true if Y is replaced by $\sum_{i=1}^{\infty} i^{-3}y_i I_{[i/N,i/N+1)}(t/N)$ for some positive integer N in which case (41) becomes

$$y_t = (t/N)^3\,e^{-(t/N)^2}\sum_{k=0}^{K}(-1)^k\left[a_k\,H_{2k}(t/N) - b_k\,H_{2k+1}(t/N)\right]. \qquad (42)$$

Computations will be more stable if N is close to the sample size T. Although in theory N must remain fixed as sample size T increases so that A and B do not drift with T, in an application one can set N to the largest value of T that one envisages arising during the course of computations.

When y_t computed according to (42) are used in connection with GMM one will need instruments to identify the $\{a_k, b_k\}_{t=0}^{K}$. The set $\{(t/N)^3e^{-(t/N)^2}H_j(t/N)\}_{j=0}^{2J+1}$ will serve as a set of identifying instruments if $J = K$ and overidentifying if $J > K$. Often models will have a scale factor multiplying y_t. This scale factor will need to be set to one to achieve identification. Similarly, an additive location parameter should be set to zero.

We end this subsection by noting that for $\Lambda_t \in \mathbb{R}^K$ the above holds with $A_K, B_K, A, B, a_k, b_k \in \mathbb{R}^K$, which amounts to applying the results above to Λ_t elementwise. Apologies for using K to mean both the dimension of Λ and the order of the sieve.

5.2 Construction of a approximate likelihood

As above, let $h : \mathbb{R}^J \times \mathbb{R}^K \times \mathbb{R}^D \to \mathbb{R}^M$ be a set of M moment conditions involving J observable variables X_t and K latent variables Λ_t. Similarly to the foregoing, a limited information state space model is defined as

$$E\left[h(X_{t+1}, \Lambda_{t+1}, \beta_0) | \mathcal{I}_t\right] = 0 \tag{43}$$

$$\Lambda_{t+1} \sim p(\Lambda_{t+1} | \mathcal{I}_{0,t}, \theta) \tag{44}$$

where $\mathcal{I}_t = \{X_1, \Lambda_1, X_2, \Lambda_2, \ldots, X_t, \Lambda_t\}$ is the information set at time t and $\mathcal{I}_{0,t} = \{\Lambda_1, \Lambda_2, \ldots, \Lambda_t\}$. The system in (43) and (44) implies a set of M conditional moment conditions

$$E[g(X_{t+1}, \Lambda_{t+1}, \theta_0) | \mathcal{I}_{0,T}] = 0. \tag{45}$$

Estimation of $\theta_0 = (\beta_0, A_0, B_0, \Sigma_0)$ is the objective of the statistical analysis, where (A_0, B_0) is the representation of $\{\mathcal{I}_{0,t} : t = 1, \ldots, \infty\}$ given in Section 5.1.

The moment equations $g(X_{t+1}, \Lambda_{t+1}, \theta)$ have exactly the same form as would obtain in a conventional GMM analysis were Λ_{t+1} observed from data. The difference is that instead of Λ_{t+1} being observed it is computed using (38) from the successive choices for (A, B), as approximated by (A_K, B_K), chosen during the course of computations by an optimization algorithm. With this, the notation above becomes redundant because knowledge of θ, hence (A, B), implies knowledge of $\{\Lambda_1, \Lambda_2, \ldots, \Lambda_T\}$. Therefore, hereafter we write

$$g(X_{t+1}, \theta) \tag{46}$$

instead of $g(X_{t+1}, \Lambda_{t+1}, \theta)$.

The approximate likelihood $\widehat{p}(X, \theta)$ is based on a transformation of the GMM objective function:

$$\widehat{p}(X, \theta) = [2\pi \det \Sigma(X, \theta)]^{-M/2} \exp\left\{-\frac{1}{2}\, g_T(X, \theta)'\, [\Sigma(X, \theta)]^{-1}\, g_T(X, \theta)\right\}. \tag{47}$$

where

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

$$
\begin{aligned}
\Sigma(X,\theta) &= \frac{1}{T}\sum_{t=1}^{T}\left[\widetilde{g}(X_t,\theta)\right]\left[\widetilde{g}(X_t,\theta)\right]', \qquad (48)\\
\widetilde{g}(X_t,\theta) &= g(X_t,\theta) - \frac{1}{\sqrt{T}}g_T(X,\theta).
\end{aligned}
$$

Depending on the nature of the dynamics and moments, a HAC estimator (Gallant, 1987, p. 445) of $\Sigma(X,\Lambda,\theta)$ may have to be substituted for (48). Given the approximate likelihood $\widehat{p}(X,\theta)$, we propose solving

$$\max_{\theta\in\Theta}\widehat{p}(X,\theta) \qquad (49)$$

or drawing from a posterior distribution

$$\widehat{p}(\theta|X) \propto \widehat{p}(X,\theta)p(\theta), \qquad (50)$$

either of which can be accomplished by MCMC. The output of the procedure is a Markov chain that can be used to construct estimators of θ_0. Another attractive feature of the procedure is that it recovers a limited information distribution of the latent variables Λ conditional on X. In many applications, this is useful as it may serve as a way to assess the quality of the model.

5.3 Theoretical justification

Direct application of Chen and Shen (1998) and Chen, Liao, and Sun (2012).

5.4 Example

We applied the particle filter method and the sieve method to the same realization of the stochastic volatility model of Section 2.1 with ρ set to zero. Comparing Figures 13 and 14 one sees that the sieve estimate of Λ appears overly smooth and far less accurate than the particle filter estimate.

(Figure 13 about here)

(Figure 14 about here)

6 Conclusion

We have presented three methods for estimating the parameters of dynamic models with unobserved variables using only moment conditions.

The first, Bayes, has a well developed theory and performs well in applications. It is computationally intensive.

The second, particle filtering, has a less well developed theory but does perform well in the computational experiments that we have undertaken. It is computationally intensive.

The third, sieves, has a well developed theory but performs erratically in the computational experiments that we have undertaken. It appears to markedly over smooth and sometimes misses the history of the latent variables entirely. It also has several tuning parameters that need adjustment, which can be time consuming. The problems just noted may be the consequence of a poor choice of a sieve to implement the method. The sieve approach is far less computationally demanding than the other two methods.

7 References

Ackerberg, Daniel, John Geweke, and Jinyong Hahn (2009), "Convergence Properties of the Likelihood of Computed Dynamic Models," *Econometrica* 77, 2009–2017.

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.

Bansal, R., and A. Yaron (2004), "Risks For the Long Run: A Potential Resolution of Asset Pricing Puzzles." *Journal of Finance 59,* 1481–1509.

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

Chen, Xiohong, and Xiotong Shen (1998), "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica* 66, 289–314.

Chen, Xiohong, Zhipeng Liao, and Xixiao Sun (2012), "Sieve Inference on Semi-Nonparametric Time Series Models," Cowles Foundation discussion Paper No. 1849.

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),” *Journald of Monetary Economics,* 55, 1191–1208.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977), “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society. Series B* 39, 1-38.

Duffie, D. and K. J. Singleton (1993), “Simulated moments estimation of Markov models of asset prices,” *Econometrica,* 61, 929–952.

Fama, E., and K. French (1993), “Common Risk Factors in the Returns on Stocks and Bonds.” *Journal of Financial Economics 33,* 3–56.

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.

Gallant, A. R., and D. W. Nychka (1987), "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica 55,* 363–390.

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

Jacquier, E., Polson, N., and Rossi, P. (2004), A Bayesian analysis fat-tailed stochastic volatility models with correlated errors. *Journal of Econometrics* 122, 185–212.

Pitt, Michael K. (2002), "Smooth Particle Filters for Likelihood Evaluation and Maximization," Working paper, The University of Warwick, UK.

Rahman, Matiur (2011), *Applications of Fourier Transforms to Generalized Functions*, MIT Press.

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

Table 1. EM PF GMM Estimates for the SV Model

Parameter	True Value	Mean	Mode	Standard Error
	Jacobian Term in Psudo-Likelihood			
ρ	0.9	0.94632	0.94204	0.029038
ϕ	0.9	0.92683	0.94940	0.028233
σ	0.5	0.43177	0.37377	0.038772
	No Jacobian Term in Psudo-Likelihood			
ρ	0.9	0.92717	0.96607	0.044129
ϕ	0.9	0.85233	0.98538	0.133490
σ	0.5	0.32223	0.52055	0.287690
	Flury and Shephard Estimator			
ρ	0.9	0.91283	0.91525	0.008078
ϕ	0.9	0.85020	0.83711	0.051060
σ	0.5	0.72344	0.77463	0.104650

Data of length $T = 100$ was generated by simulating the model of Subsection 2.1 at the values shown in the column labeled true. In the first two panels the model was estimated by using the methods described in Section 4. In the third panel the estimator is the Bayesian estimator proposed by Flury and Shepard (2010). 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. In all panels the number of particles is $N = 5000$. The columns labeled mean, mode, and standard deviation are the mean, mode, and standard deviations of an MCMC chain of length 400,000.

Table 2. EM PF GMM Estimates for the DSGE Model

Parameter	True Value	Mean	Mode	Standard Error
		Jacobian Term in Psudo-Likelihood		
ρ_z	0.15	0.14848	0.14654	0.00154
ρ_ϕ	0.68	0.66901	0.67450	0.00547
ϕ_λ	0.56	0.56005	0.55316	0.00561
σ_z	0.71	0.70081	0.70593	0.00592
σ_ϕ	2.93	2.92730	2.92010	0.00481
σ_λ	0.11	0.11096	0.11629	0.00354
ν	0.96	0.96017	0.96305	0.00334
β	0.996	0.996	0.996	fixed
		No Jacobian Term in Psudo-Likelihood		
ρ_z	0.15	0.23546	0.05580	0.11487
ρ_ϕ	0.68	0.56419	0.64537	0.10375
ϕ_λ	0.56	0.65883	0.63018	0.05018
σ_z	0.71	0.57542	0.53044	0.11520
σ_ϕ	2.93	2.88670	2.84500	0.13054
σ_λ	0.11	0.11661	0.11726	0.00799
ν	0.96	0.85246	0.84432	0.09713
β	0.996	0.996	0.996	fixed

Data of length $T = 100$ was generated by simulating the model of Subsection 2.2 at the values shown in the column labeled true. The model was estimated by using the methods described in Section 4 with a Jacobian term. The number of particles is $N = 10000$. The columns labeled mean, mode, and standard deviation are the mean, mode, and standard deviations of an MCMC chain of length 200000.

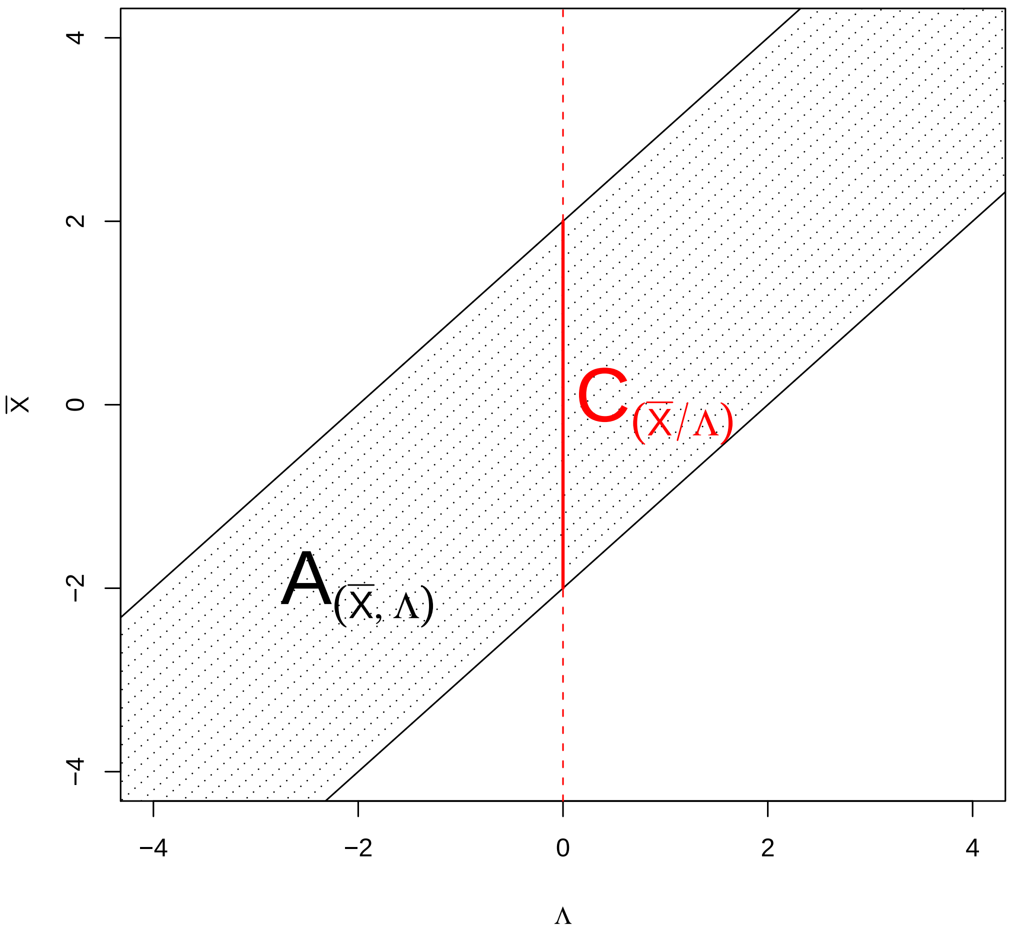


Figure 1. GMM Probability Assignment. Under the assumption that $\bar{X}$ and Λ have joint density $p(\bar{X},\Lambda) = \frac{1}{\sqrt{2\pi}}e^{-\frac{n}{2}\frac{(\bar{X}-\Lambda)^2}{s^2}}$ where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, joint probability on $(\bar{X},\Lambda)$ can only be assigned to sets bounded by 45 degree lines such as the one labeled $B_{(\bar{X},\Lambda)}$ 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.

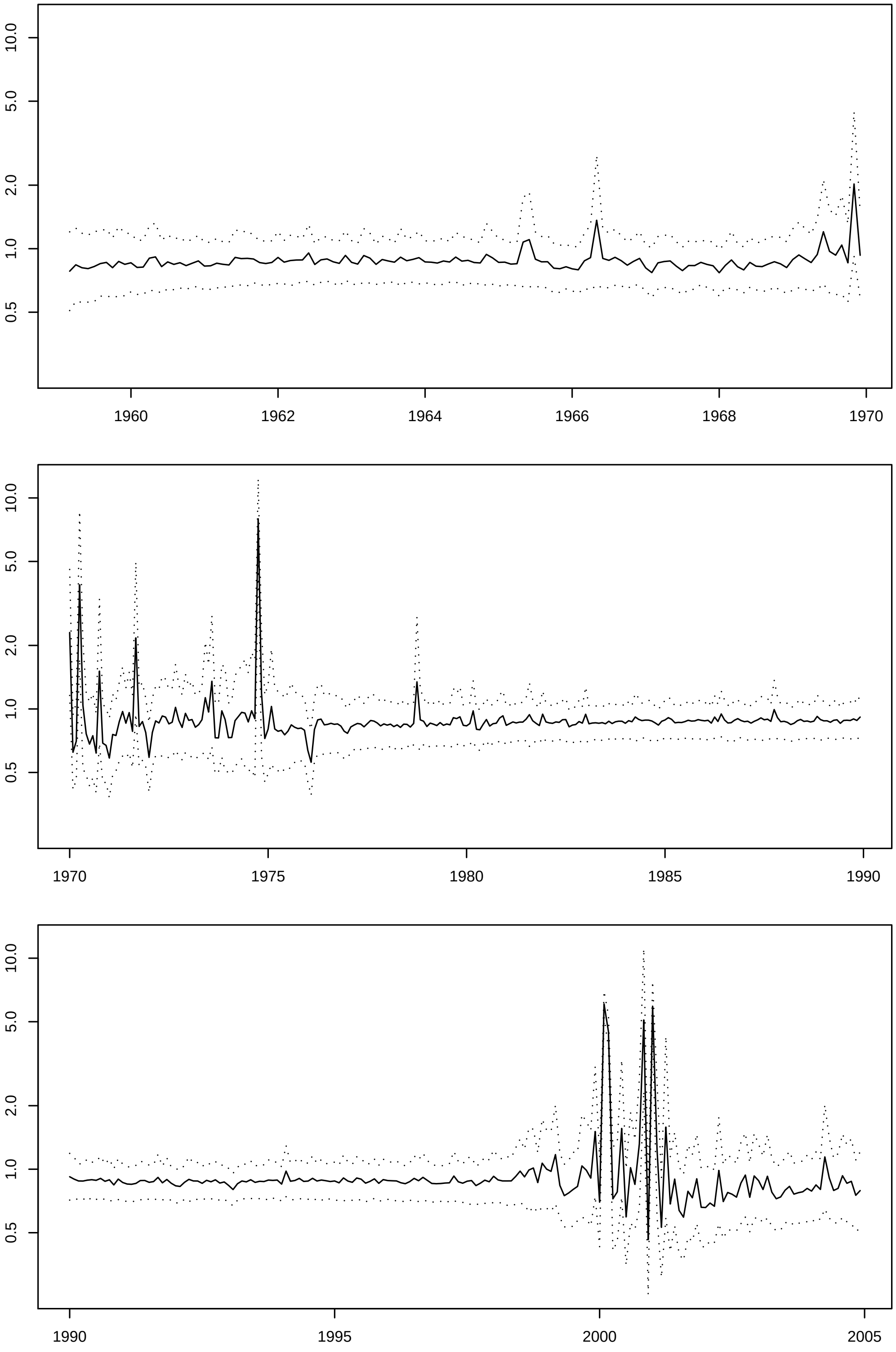


Figure 2. The Posterior Mean of the Monthly Pricing Kernel. Plotted as the solid line is the posterior mean of $\log(\Lambda_2), ..., \log(\Lambda_{551})$ under a loose prior. The dotted lines are plus and minus one standard deviation. The units of the vertical axis are the exponential of the plotted quantity.

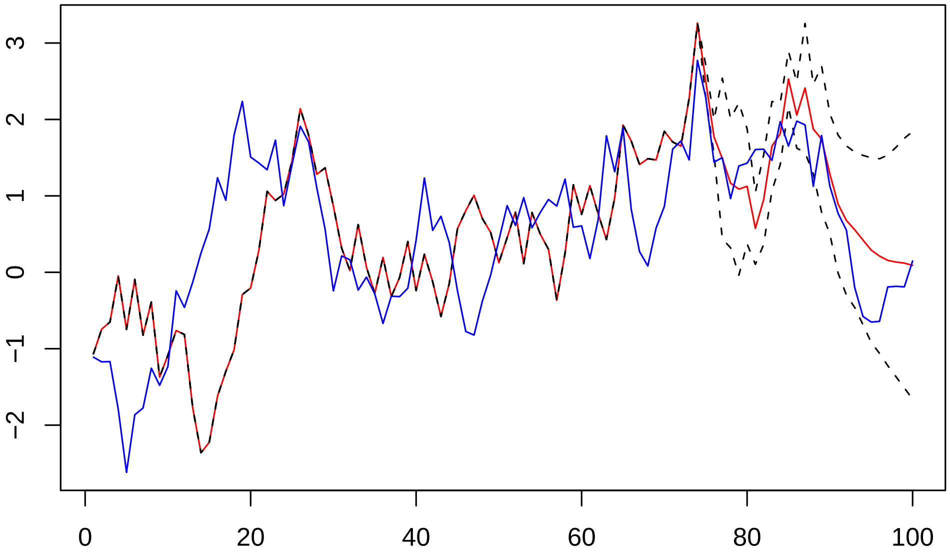


Figure 3. Particle Filter Estimate of Λ, Time Series Plot. Data of length $T = 100$ was generated from a simulation of the model of Subsection 2.1 and estimated using the method described in Section 4 with a Jacobian term. The blue line plots the simulated Λ. At the values in the column labeled True in Table 1, the red line is the pointwise mean of the $N = 5000$ particles and the dashed black lines are plus and minus two pointwise standard errors. The moment equations were (1) through (6); a one lag HAC estimator was used for (24).

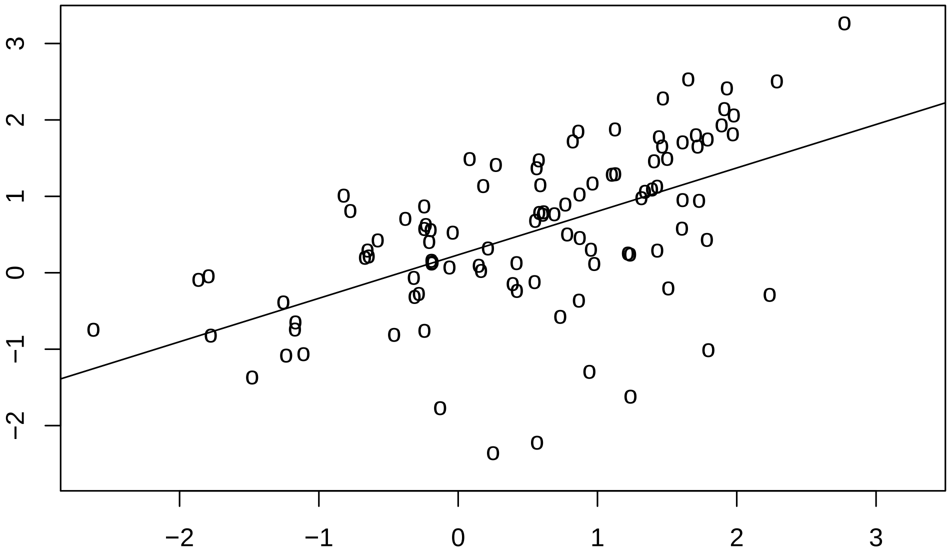


Figure 4. Particle Filter Estimate of Λ, Scatter Plot. As for Figure 3 except that plotted is the pointwise mean of the $N = 5000$ particles vs. the simulated Λ.

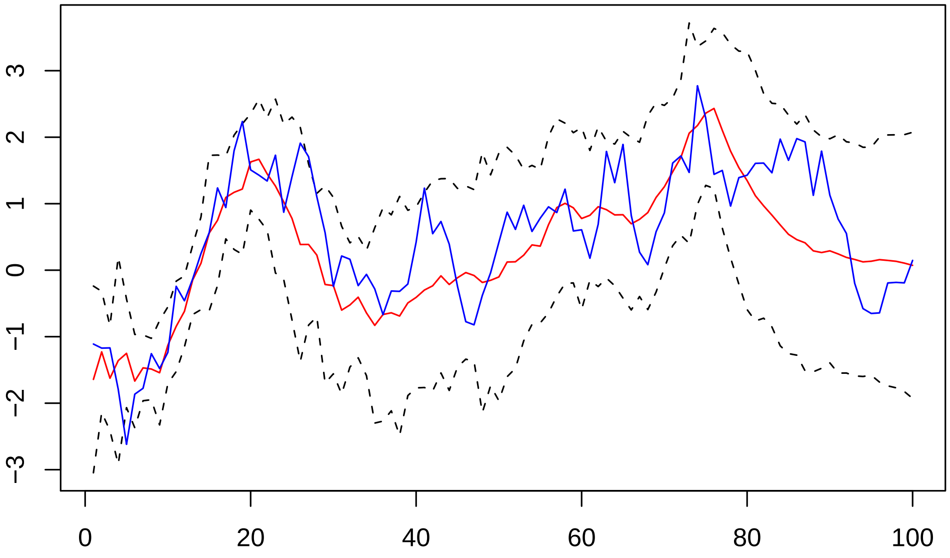


Figure 5. Particle Filter Estimate of Λ, Time Series Plot. As for Figure 3 except that estimation is without a Jacobian term.

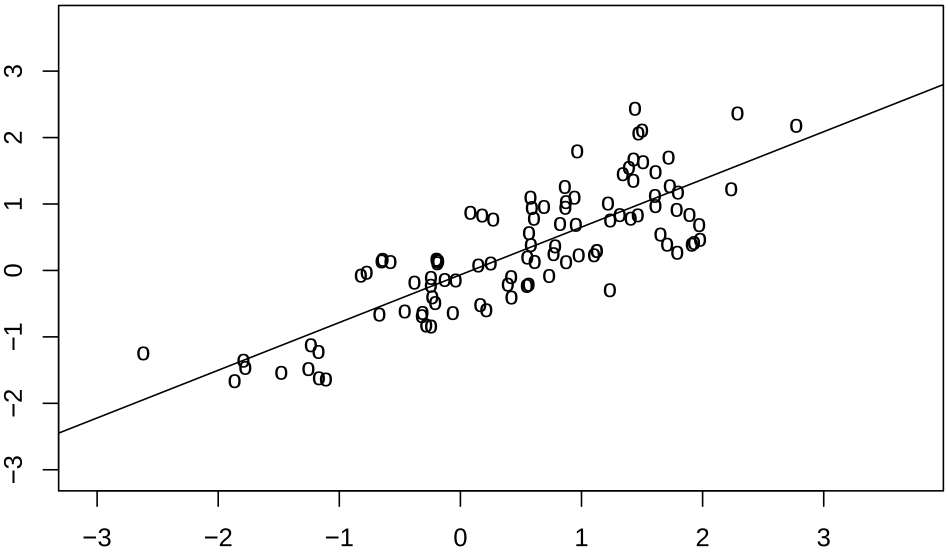


Figure 6. Particle Filter Estimate of Λ, Scatter Plot. As for Figure 5 except that plotted is the pointwise mean of the $N = 5000$ particles vs. the simulated Λ.

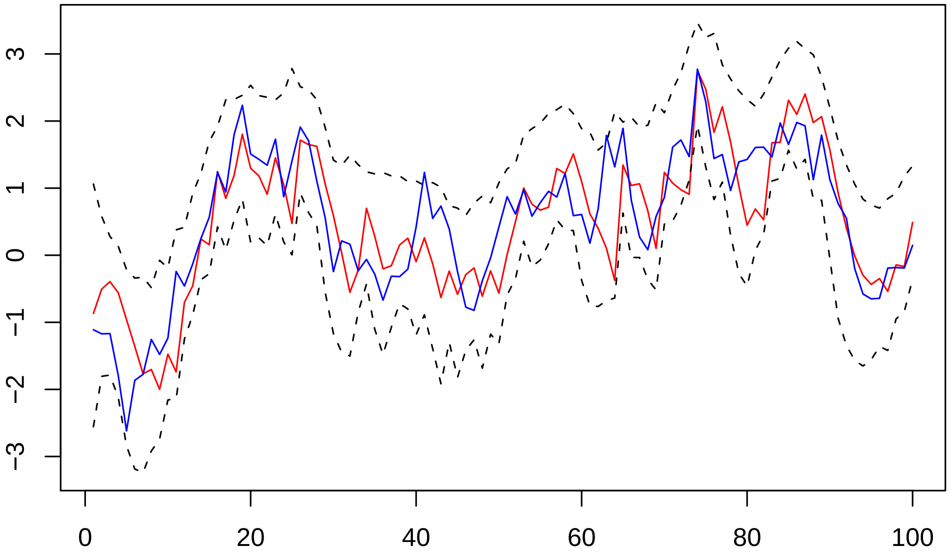


Figure 7. Particle Filter Estimate of Λ, Time Series Plot. As for Figure 3 except that plotted is a filter, not a smooth, and weighting is by the measurement density, not GMM.

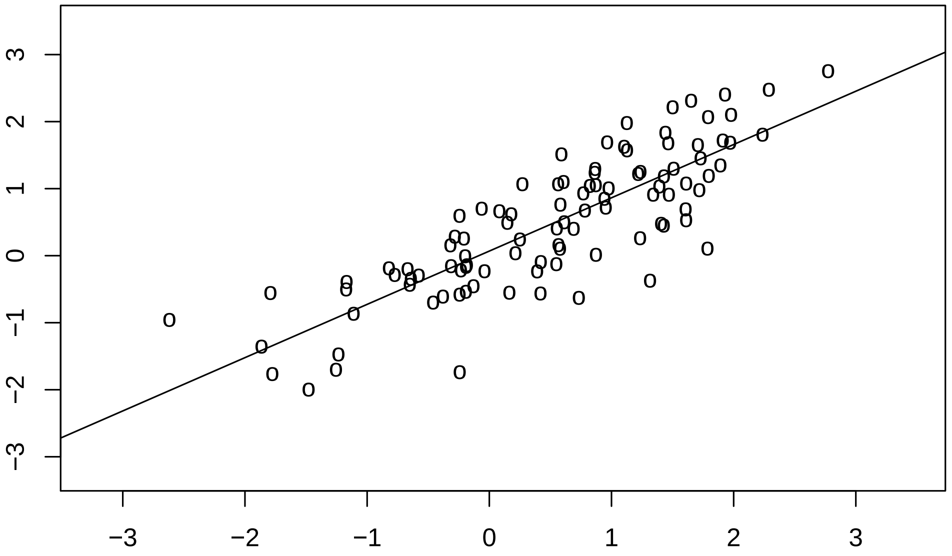
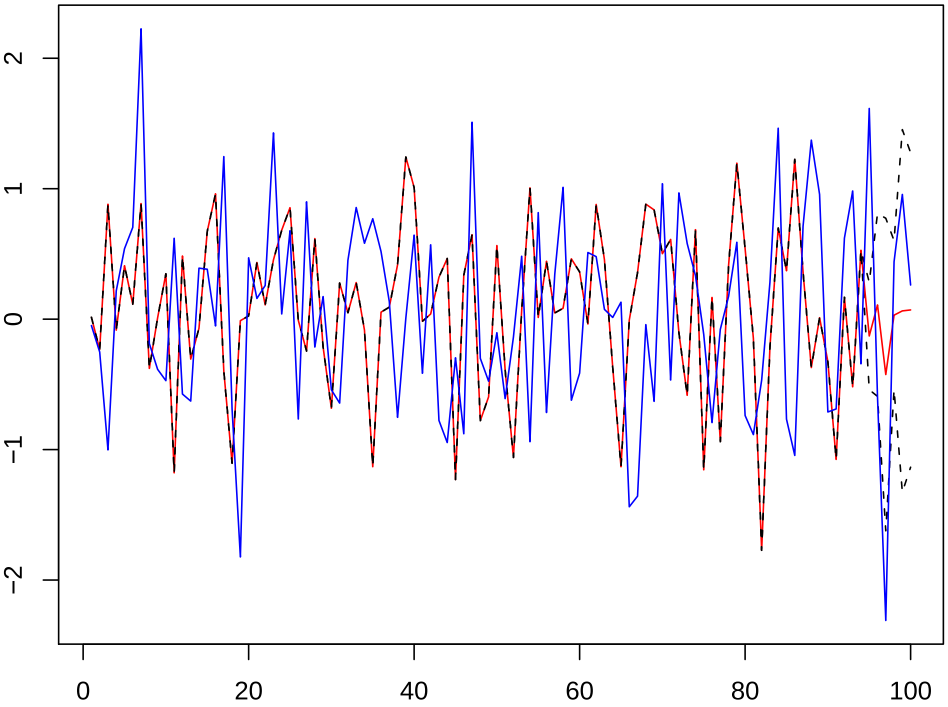


Figure 8. Particle Filter Estimate of Λ**, Scatter Plot.** As for Figure 7 except that plotted is the pointwise mean of the $N = 5000$ particles vs. the simulated Λ.



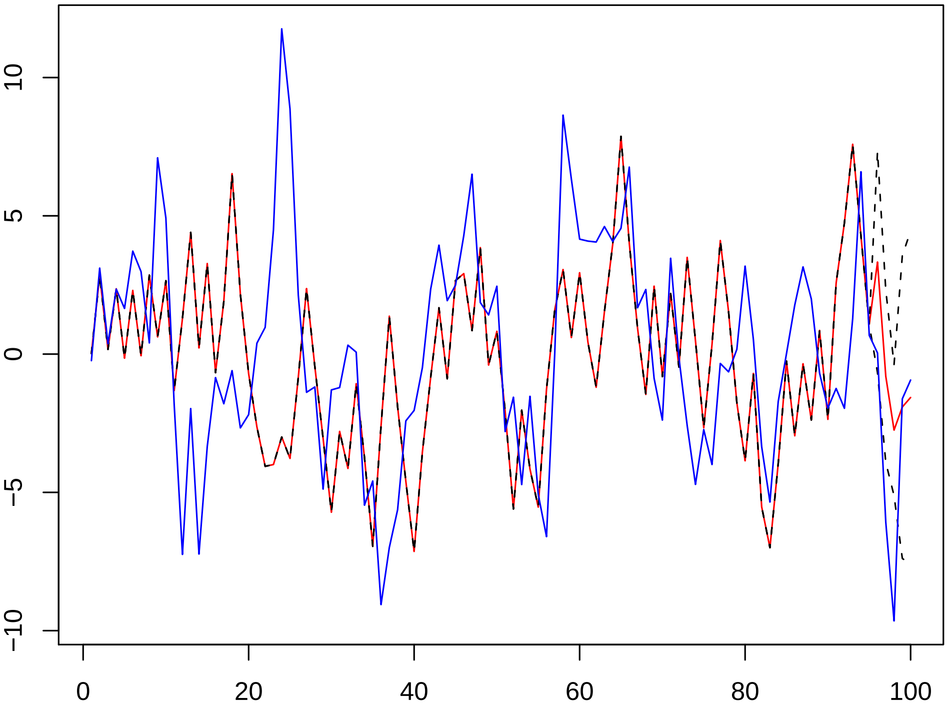


Figure 9. PF Estimate of Λ with Jacobian, Time Series Plot. Data of length $T = 100$ was generated by simulating the model of Section 2. The blue line plots the simulated Λ. at the values in the column labeled True in Table 2. The red line is the pointwise mean of the $N = 10000$ particles and the dashed black lines are plus and minus two pointwise standard errors. The moment equations were (7) through (14); a two lag HAC estimator was used for (24).

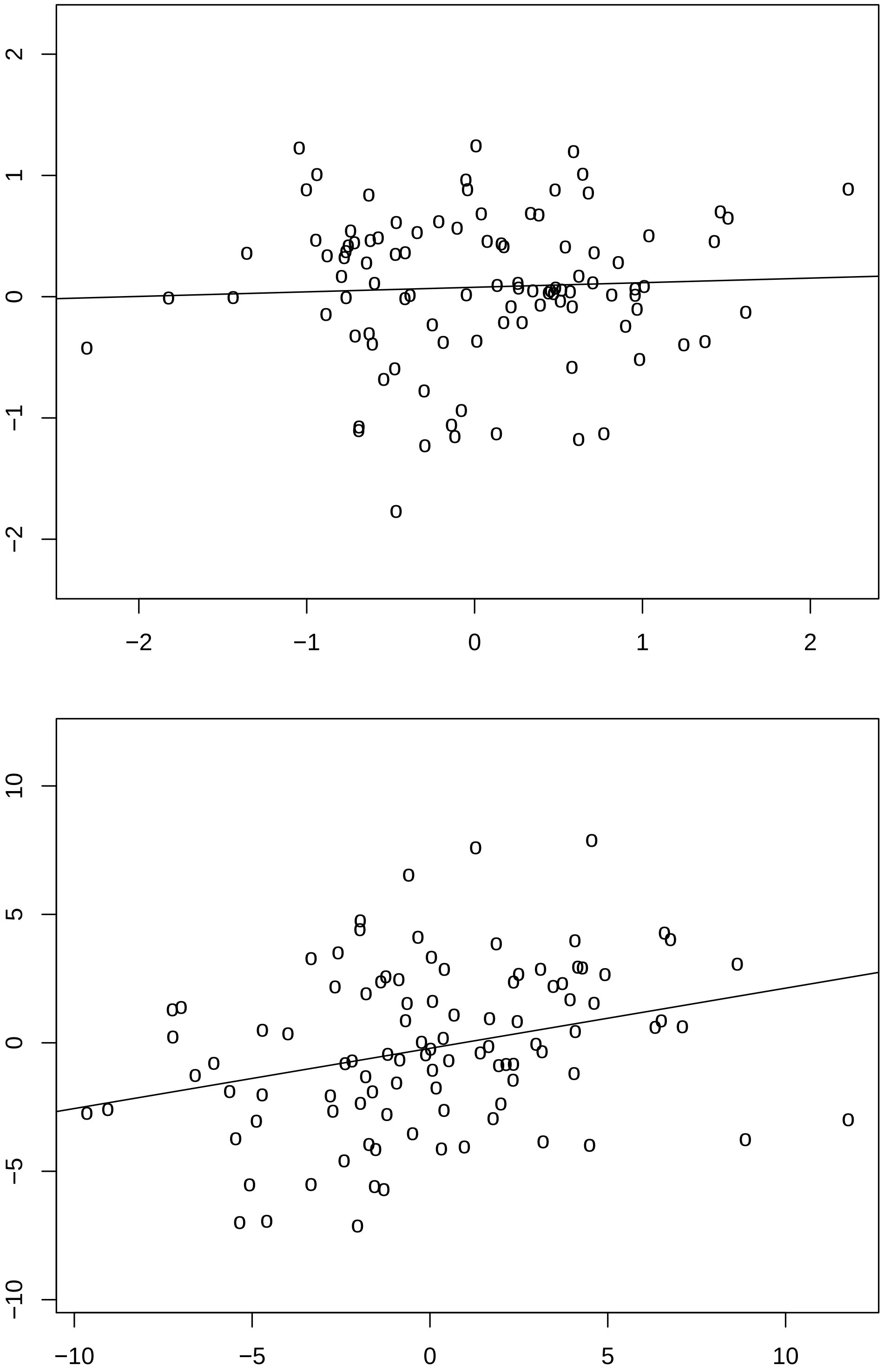


Figure 10. As for Figure 9 except that plotted is the pointwise mean of the $N = 10000$ particles vs. the simulated Λ.

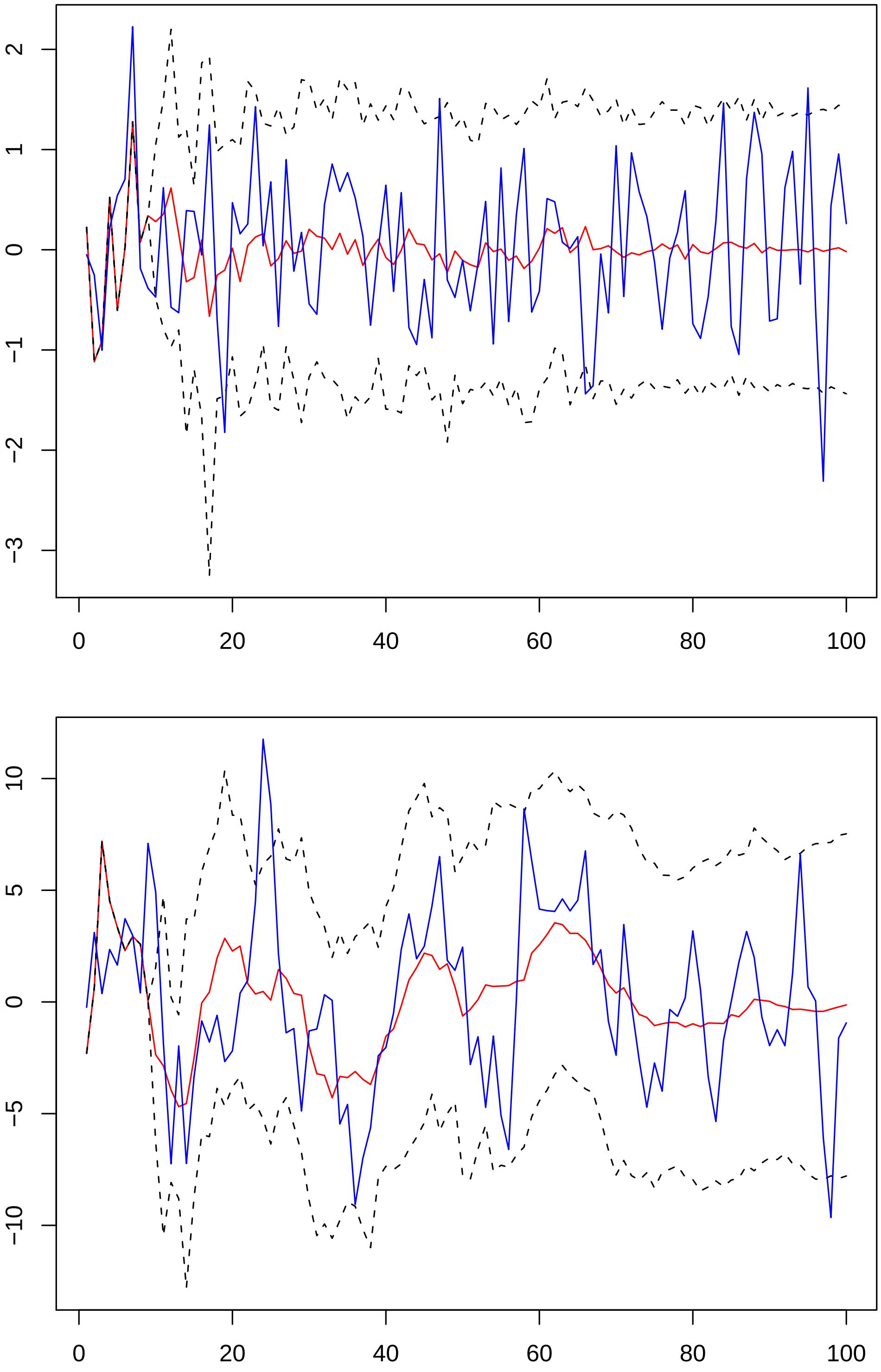


Figure 11. PF Estimate of Λ without Jacobian, Time Series Plot. As for Figure 9 but without a Jacobian term in the psuedo likelihood

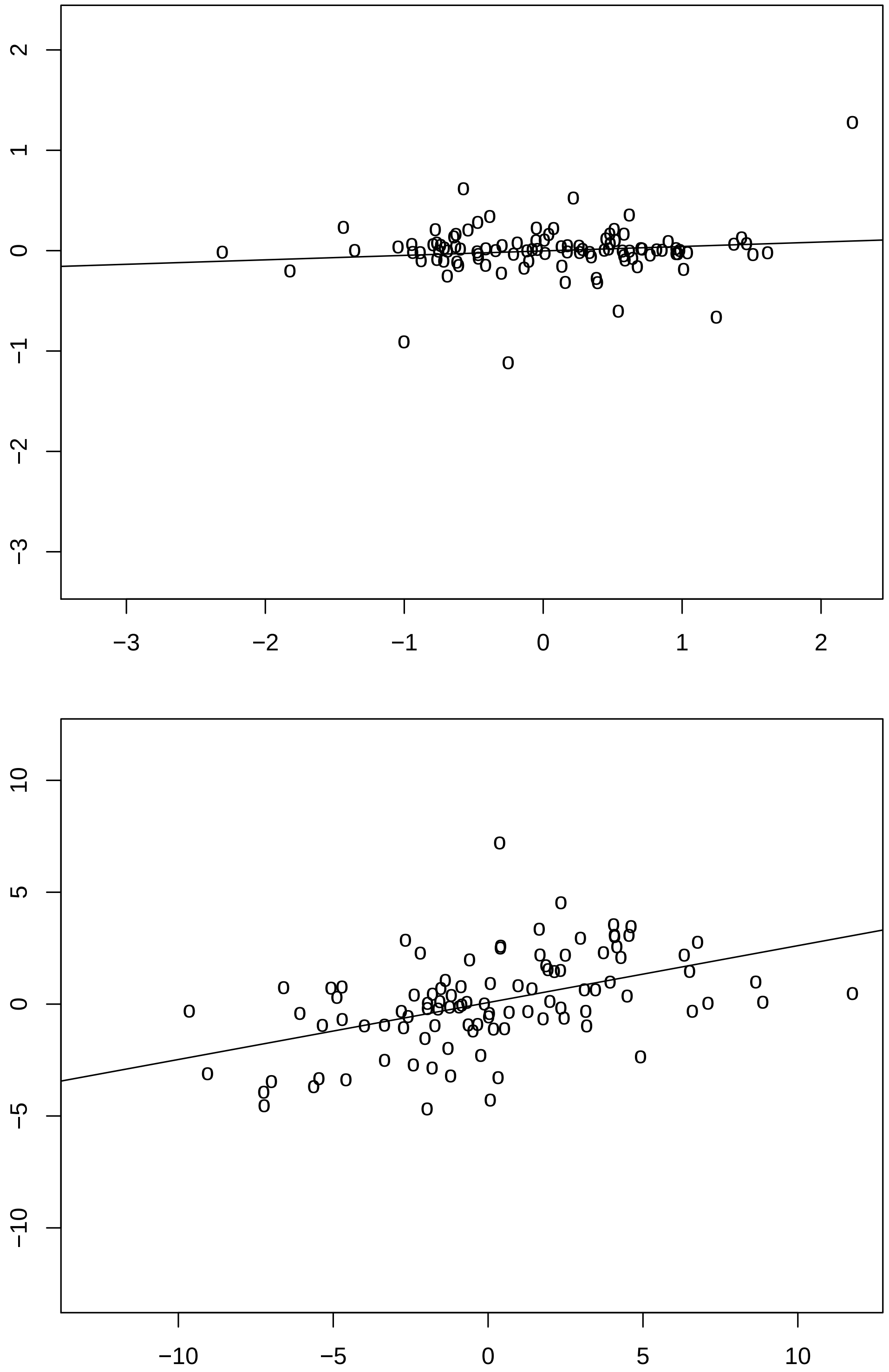


Figure 12. PF Estimate of Λ without Jacobian, Scatter Plot. As for Figure 10 but without a Jacobian term in the psuedo likelihood

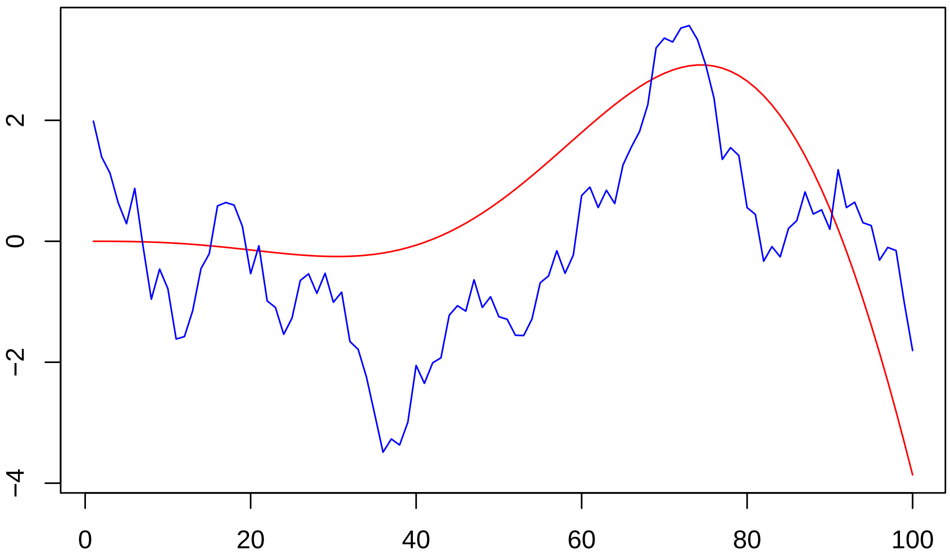


Figure 13. Sieve Estimate of Λ. Data of length $T = 100$ was generated by simulating the model of Section 2 at the parameter values shown except that $\rho = 0$. The blue line plots the simulated Λ. The red line is maximum likelihood estimate of Λ using the likelihood described in Subsection 5.2 with $K = 6$.