GMM with Latent Variables

A. Ronald Gallant
Penn State University

Raffaella Giacomini
University College London

Giuseppe Ragusa
Luiss University
Contribution

• The contribution of GMM (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.
The Requirements, 1 of 3

- A structural model with parameters $\theta$ and true value $\theta^o$

- Observed variables: $X = (X_1, X_2, ..., X_T)$

- Latent variables: $\Lambda = (\Lambda_1, \Lambda_2, ..., \Lambda_T)$

- Known transition density: $\Lambda_{t+1} \sim P(\Lambda_{t+1} | \Lambda_t, \theta)$

- Conditional moment conditions: $\mathcal{E}[g(X_{t+1}, \Lambda_{t+1}, \theta) | \mathcal{I}_t] = 0$
  - That would identify $\theta$ if both $X$ and $\Lambda$ were observed.
  - $\mathcal{I}_t = \{X_{-\infty}, ..., X_t, \Lambda_{-\infty}, ..., \Lambda_t\}$
The Requirements, 2 of 3

- Sample moment conditions

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

- Weighting matrix (May have to use a HAC weighting matrix instead.)

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

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

- Such that

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

- Hansen and Singleton (1982)
- Gallant and White (1987)
A sample \( \{\theta^{(i)}\}_{i=1}^{\mathcal{R}} \) from the density

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

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

— Chernozhukov and Hong (2003)
Estimation Strategy

- Sample \( \{\theta^{(i)}, \Lambda^{(i)}\} \) from the density

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

- Might multiply \( p(\theta, \Lambda) \) by a Jacobian term \( \det \Sigma(X, \Lambda, \theta)^{-M/2} \)

- Metropolis within Gibbs algorithm
  - Sample \( \theta^{(i)} \) given \( \Lambda^{(i-1)} \) and \( \theta^{(i-1)} \) using Metropolis
    * last draw of MCMC chain of length \( K \).
  - Sample \( \Lambda^{(i)} \) given \( \theta^{(i)} \) and \( \Lambda^{(i-1)} \) using Gibbs.
    * last particle of a modified particle filter of size \( N \).
  - Iterate back and forth. (Can view it as an approximate EM algorithm.)

- Estimate and scale are mean and standard deviation of \( \{\theta^{(i)}\} \).
Next:
Two Examples

- A Dynamic Stochastic General Equilibrium Model
  - Description
  - Estimates

- A Stochastic Volatility Model
  - Description
  - Estimates
A DSGE Model – 1 of 4

From Del Negro and Schorfheide (2008) simplified to permit an analytic solution by removing rigidities, investment, etc.

Three shocks:

\[
\begin{align*}
    z_t &= \rho_z z_{t-1} + \sigma_z \epsilon_{z,t} \quad \text{Factor productivity} \\
    \phi_t &= \rho_{\phi} \phi_{t-1} + \sigma_{\phi} \epsilon_{\phi,t} \quad \text{Consumption/leisure preference} \\
    \lambda_t &= \rho_{\lambda} \lambda_{t-1} + \sigma_{\lambda} \epsilon_{\lambda,t} \quad \text{Price elasticity of intermediate goods}
\end{align*}
\]

Three outputs:

\[
\begin{align*}
    w_t & \quad \text{Wages} \\
    y_t & \quad \text{Output} \\
    \pi_t & \quad \text{Inflation}
\end{align*}
\]
A DSGE Model – 2 of 4

First order conditions

\[
0 = y_t + \frac{1}{\beta} \pi_t - \mathbb{E}_t(y_{t+1} + \pi_{t+1} + z_{t+1}) \\
0 = w_t + \lambda_t \\
0 = w_t - (1 + \nu)y_t - \phi_t
\]

where \(\nu\) is a labor supply elasticity and \(\beta\) is the discount rate.

The true values of the parameters are

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

We take \(w_t, y_t,\) and \(\pi_t\) as measured and \(z_t\) and \(\phi_t\) as latent so

\[
X_t = (w_t, y_t, \pi_t) \\
\Lambda_t = (z_t, \phi_t).
\]
A DSGE Model – 3 of 4

A set of conditions that identify the model are

\[ g_1 = (w_t - \rho_\lambda w_{t-1})^2 - \sigma_\lambda^2 \]
\[ g_2 = w_{t-1}(w_t - \rho_\lambda w_{t-1}) \]
\[ g_3 = [w_{t-1} - (1 + \nu)y_{t-1}][w_t - (1 + \nu)y_t - \rho_\phi(w_{t-1} - (1 + \nu)y_{t-1})] \]
\[ g_4 = [w_{t-1} - (1 + \nu)y_{t-1}](\phi_t - \rho_\phi \phi_{t-1}) \]
\[ g_5 = [w_t - (1 + \nu)y_t]^2 - \sigma_\phi^2 \]
\[ g_6 = w_{t-1}(y_{t-1} + \frac{1}{\beta}(\pi_{t-1} - y_t - \pi_t - \rho_z z_{t-1})) \]
\[ g_7 = y_{t-1}(y_{t-1} + \frac{1}{\beta}(\pi_{t-1} - y_t - \pi_t - \rho_z z_{t-1})) \]
\[ g_8 = \pi_{t-1}(y_{t-1} + \frac{1}{\beta}(\pi_{t-1} - y_t - \pi_t - \rho_z z_{t-1})) \]
\[ g_9 = (y_{t-1} + \frac{1}{\beta}(\pi_{t-1} - y_t - \pi_t))^2 - \frac{\rho_z^2 \sigma_z^2}{1 - \rho_z^2} \]
A DSGE Model – 4 of 4

- An analytic expression for the likelihood \( L(\theta) = p(X|\theta) \) is available for this model.

- Analysis of the likelihood shows that only one of the four parameters \( \sigma_z, \sigma_\phi, \nu, \beta \) can be identified.

- Three will have to be calibrated in order to apply frequentist methods.

- We calibrate \( \sigma_z, \sigma_\phi, \nu \) and leave \( \beta \) as the free parameter.
Table 1. Parameter Estimates, DSGE Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>Mean</th>
<th>Mode</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>With Jacobian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho_z$</td>
<td>0.15</td>
<td>0.21596</td>
<td>0.15006</td>
<td>0.08632</td>
</tr>
<tr>
<td>$\rho_\phi$</td>
<td>0.68</td>
<td>0.60098</td>
<td>0.58945</td>
<td>0.04988</td>
</tr>
<tr>
<td>$\rho_\lambda$</td>
<td>0.56</td>
<td>0.50134</td>
<td>0.46443</td>
<td>0.28818</td>
</tr>
<tr>
<td>$\sigma_\lambda$</td>
<td>0.11</td>
<td>0.10827</td>
<td>0.08923</td>
<td>0.06494</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.996</td>
<td>0.98429</td>
<td>0.99603</td>
<td>0.01476</td>
</tr>
<tr>
<td>Without Jacobian</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho_z$</td>
<td>0.15</td>
<td>0.21887</td>
<td>0.23069</td>
<td>0.09179</td>
</tr>
<tr>
<td>$\rho_\phi$</td>
<td>0.68</td>
<td>0.59967</td>
<td>0.60750</td>
<td>0.04988</td>
</tr>
<tr>
<td>$\rho_\lambda$</td>
<td>0.56</td>
<td>0.50884</td>
<td>0.31473</td>
<td>0.28981</td>
</tr>
<tr>
<td>$\sigma_\lambda$</td>
<td>0.11</td>
<td>0.10797</td>
<td>0.11613</td>
<td>0.06896</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.996</td>
<td>0.98201</td>
<td>0.99634</td>
<td>0.01834</td>
</tr>
<tr>
<td>Maximum Likelihood</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho_z$</td>
<td>0.15</td>
<td>0.15165</td>
<td>0.15087</td>
<td>0.00583</td>
</tr>
<tr>
<td>$\rho_\phi$</td>
<td>0.68</td>
<td>0.59185</td>
<td>0.59419</td>
<td>0.05044</td>
</tr>
<tr>
<td>$\rho_\lambda$</td>
<td>0.56</td>
<td>0.56207</td>
<td>0.56549</td>
<td>0.05229</td>
</tr>
<tr>
<td>$\sigma_\lambda$</td>
<td>0.11</td>
<td>0.11225</td>
<td>0.11189</td>
<td>0.00508</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.996</td>
<td>0.99640</td>
<td>0.99643</td>
<td>0.00186</td>
</tr>
</tbody>
</table>

Data with $T = 250$ simulated at true values. Gibbs particles are $N = 1000$; Metropolis draws are $K = 50$. GMM mean, mode, and standard deviation are from MCMC chains of length $R = 9637$ with stride of 1; for MLE chain $R = 500000$, stride is 5.
Remark: The Gibbs draw should evaluate the moments in the Metropolis step accurately; not necessarily approximate the history accurately.
Figure 2. PF Estimate of $\Lambda$ without Jacobian, DSGE Model
Figure 3. PF Estimate of Λ with Jacobian, DSGE Model
Figure 4. PF Estimate of Λ without Jacobian, DSGE Model
The Choice of Moments Does Matter 1 of 2

• It is possible to perform counter-factual (e.g. impulse-response) analysis using moment conditions alone.

• However, for it to work, one must do a much better job of estimating the history of the latent variables.

• To estimate latent variables, it is not necessary to identify model parameters.

• Only the latent variables need to be identified.
The Choice of Moments Does Matter 2 of 2

Moment conditions for counter-factual analysis

\[
\begin{align*}
    h_1 &= y_{t-1} + \frac{1}{\beta} \pi_{t-1} - y_t - \pi_t - \rho z z_{t-1} \\
    h_2 &= w_{t-1} h_1 \\
    h_3 &= y_{t-1} h_1 \\
    h_4 &= \pi_{t-1} h_1 \\
    h_5 &= w_t - (1 + \nu) y_t - \phi_t \\
    h_6 &= w_{t-1} h_5 \\
    h_7 &= y_{t-1} h_5 \\
    h_8 &= \pi_{t-1} h_5
\end{align*}
\]
Figure 5. PF Estimate of Λ with Jacobian, DSGE Model
Figure 6. PF Estimate of $\Lambda$ without Jacobian, DSGE Model
Figure 7. PF Estimate of $\Lambda$ with Jacobian, DSGE Model
Figure 8. PF Estimate of \( \Lambda \) without Jacobian, DSGE Model
Gibbs and Metropolis Moments Can Differ

If we use the moments that identify the model used for Table 1 for the Metropolis step and the moments designed for a counterfactual analysis used for Figures 5 through 8 for the Gibbs step, we get slightly better results in the following Table 2.
Table 2. Alternative Parameter Estimates, DSGE Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>Mean</th>
<th>Mode</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>With Jacobian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho_z$</td>
<td>0.15</td>
<td>0.21702</td>
<td>0.15006</td>
<td>0.08367</td>
</tr>
<tr>
<td>$\rho_\phi$</td>
<td>0.68</td>
<td>0.61408</td>
<td>0.58945</td>
<td>0.05102</td>
</tr>
<tr>
<td>$\rho_\lambda$</td>
<td>0.56</td>
<td>0.50082</td>
<td>0.46443</td>
<td>0.28344</td>
</tr>
<tr>
<td>$\sigma_\lambda$</td>
<td>0.11</td>
<td>0.11086</td>
<td>0.08924</td>
<td>0.06493</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.996</td>
<td>0.98740</td>
<td>0.99603</td>
<td>0.01056</td>
</tr>
<tr>
<td></td>
<td>Without Jacobian</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho_z$</td>
<td>0.15</td>
<td>0.23508</td>
<td>0.15007</td>
<td>0.08975</td>
</tr>
<tr>
<td>$\rho_\phi$</td>
<td>0.68</td>
<td>0.69870</td>
<td>0.58945</td>
<td>0.06127</td>
</tr>
<tr>
<td>$\rho_\lambda$</td>
<td>0.56</td>
<td>0.49904</td>
<td>0.46443</td>
<td>0.28418</td>
</tr>
<tr>
<td>$\sigma_\lambda$</td>
<td>0.11</td>
<td>0.11292</td>
<td>0.08924</td>
<td>0.06559</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.996</td>
<td>0.97465</td>
<td>0.99604</td>
<td>0.02479</td>
</tr>
<tr>
<td></td>
<td>Maximum Likelihood</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho_z$</td>
<td>0.15</td>
<td>0.15165</td>
<td>0.15087</td>
<td>0.00583</td>
</tr>
<tr>
<td>$\rho_\phi$</td>
<td>0.68</td>
<td>0.59185</td>
<td>0.59419</td>
<td>0.05044</td>
</tr>
<tr>
<td>$\rho_\lambda$</td>
<td>0.56</td>
<td>0.56207</td>
<td>0.56549</td>
<td>0.05229</td>
</tr>
<tr>
<td>$\sigma_\lambda$</td>
<td>0.11</td>
<td>0.11225</td>
<td>0.11189</td>
<td>0.00508</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.996</td>
<td>0.99640</td>
<td>0.99643</td>
<td>0.00186</td>
</tr>
</tbody>
</table>

Data with $T = 250$ simulated at true values. Gibbs particles are $N = 1000$; Metropolis draws are $K = 50$. GMM mean, mode, and standard deviation are from MCMC chains of length $R = 9637$ with stride of 1; for MLE chain $R = 500000$, stride is 5.
A Stochastic Volatility Model – 1 of 2

\[ X_t = \rho X_{t-1} + \exp(\Lambda_t) u_t \] (1)
\[ \Lambda_t = \phi \Lambda_{t-1} + \sigma e_t \] (2)
\[ e_t \sim N(0, 1) \] (3)
\[ u_t \sim N(0, 1) \] (4)

The true values of the parameters are

\[ \theta_0 = (\rho_0, \phi_0, \sigma_0) = (0.9, 0.9, 0.5) \] (plots)
\[ \theta_0 = (\rho_0, \phi_0, \sigma_0) = (0.25, 0.8, 0.1) \] (estimation)
A Stochastic Volatility Model – 2 of 2

Moment Conditions

\[ h_1 = (X_t - \rho X_t)^2 - [\exp(\Lambda_t)]^2 \]
\[ h_2 = |X_t - \rho X_t||X_{t-1} - \rho X_{t-1}| - \left(\frac{2}{\pi}\right)^2 \exp(\Lambda_t) \exp(\Lambda_{t-1}) \]
\[ \vdots \]
\[ h_{L+1} = |X_t - \rho X_t||X_{t-L} - \rho X_{t-L}| - \left(\frac{2}{\pi}\right)^2 \exp(\Lambda_t) \exp(\Lambda_{t-L}) \]
\[ h_{L+2} = X_{t-1}(X_t - \rho X_{t-1}) \]
\[ h_{L+3} = \Lambda_{t-1}(\Lambda_t - \phi \Lambda_{t-1}) \]
\[ h_{L+4} = (\Lambda_t - \phi \Lambda_{t-1})^2 - \sigma^2 \]
Table 3. Parameter Estimates, SV Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>Mean</th>
<th>Mode</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>With Jacobian Term</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.25</td>
<td>0.30488</td>
<td>0.30961</td>
<td>0.074778</td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.8</td>
<td>0.09153</td>
<td>0.94851</td>
<td>0.660790</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>0.1</td>
<td>0.09023</td>
<td>0.06702</td>
<td>0.050229</td>
</tr>
</tbody>
</table>

|           | Without Jacobian |      |      |                |
| $\rho$    | 0.25       | 0.30271 | 0.30939 | 0.076758      |
| $\phi$    | 0.8        | 0.15348 | 0.85765 | 0.643400      |
| $\sigma$  | 0.1        | 0.11400 | 0.08435 | 0.070081      |

|           | Flury and Shephard Estimator |      |      |                |
| $\rho$    | 0.25       | 0.30278 | 0.28555 | 0.059320      |
| $\phi$    | 0.8        | 0.17599 | 0.89189 | 0.509780      |
| $\sigma$  | 0.1        | 0.09737 | 0.07839 | 0.064661      |

Data of length $T = 200$ was generated from the SV model at true values. In all panels the number of particles is $N = 1000$. The columns labeled mean, mode, and standard deviation are the mean, mode, and standard deviations of an MCMC chain of length 200000.
Figure 9. PF Estimate of $\Lambda$ with Jacobian Term, SV Model
Figure 10. PF Estimate of $\Lambda$ without Jacobian, SV Model
Figure 11. Flury-Shephard Estimate of \( \Lambda \), SV Model
Figure 12. PF Estimate of Λ with Jacobian Term, SV Model
Figure 13. PF Estimate of Λ without Jacobian, SV Model
Figure 14. Flury-Shephard Estimate of $\Lambda$, SV Model
Next:
The Three Algorithms

• A particle filter algorithm
  – Input: $\theta$
  – Output: Draws $\{\Lambda^{(i)}\}_{i=1}^R$ from $P(\Lambda | X, \theta)$

• Gibbs algorithm
  – Input: Draws $\theta^{(i-1)}$ and $\Lambda^{(i-1)}$
  – Output: A draw $\Lambda^{(i)}$ from $P(\Lambda | X, \theta)$

• Metropolis algorithm
  – Input: Draws $\theta^{(i-1)}$ and $\Lambda^{(i)}$
  – Output: A draw $\theta^{(i)}$ from $P(\theta | X, \Lambda)$
Notation

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

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

- $p(X_{1:t}, \Lambda_{1:t}, \theta) = (2\pi)^{-M/2} \exp\left\{-\frac{1}{2}g_t(X_{1:t}, \Lambda_{1:t}, \theta)' [\Sigma(X_{1:t}, \Lambda_{1:t}, \theta)]^{-1} g_t(X_{1:t}, \Lambda_{1:t}, \theta)\right\}$
Particle Filter Algorithm, 1 of 3

1. Initialization.

- Input $\theta$ (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 $p(\Lambda_t|\Lambda_{t-1}, \theta)$.

- Set $t$ to $T_0 + 1$.

- Set $\Lambda_{1:t-1}^{(i)} = (\Lambda_1^{(i)}, \Lambda_2^{(i)}, \ldots, \Lambda_{T_0}^{(i)})$
Particle Filter Algorithm, 2 of 3

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 \( \tilde{w}_t^{(i)} = p(X_{1:t}, \tilde{\Lambda}_{1:t}^{(i)}, \theta) \).

• Scale the weights to sum to one.

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

4. Repeat

• If $t < T$, increment $t$ and go to Importance Sampling Step;

• else output $\left\{ \Lambda^{(i)}_{1:T} \right\}_{i=1}^{N}$. 
Gibbs Algorithm, 1 of 3

1. Initialization.

- Input $\Lambda_{1:T}^{(1)}$, $\theta$ (and $X$).

- Set $T_0$ to the minimum sample size required to compute $g_t(X_{1:t}, \Lambda_{1:t}, \theta)$.

- For $i = 2, \ldots, N$ sample $(\Lambda_1^{(i)}, \Lambda_2^{(i)}, \ldots, \Lambda_{T_0}^{(i)})$ from $p(\Lambda_t|\Lambda_{t-1}, \theta)$.

- Set $t$ to $T_0 + 1$.

- Set $\Lambda_{1:t-1}^{(i)} = (\Lambda_1^{(i)}, \Lambda_2^{(i)}, \ldots, \Lambda_{T_0}^{(i)})$. 
Gibbs Algorithm, 2 of 3

2. Importance sampling step.

- For $i = 2, \ldots, N$ 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 $i = 1, \ldots, N$ compute weights $\tilde{w}_t^{(i)} = p(X_{1:t}, \tilde{\Lambda}_{1:t}^{(i)}, \theta)$.

- Scale the weights to sum to one.

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

4. Repeat

- If \( t < T \), increment \( t \) and go to Importance Sampling Step;

- else output the particle \( \Lambda_{1:T}^{(N)} \).
Metropolis Algorithm

Proposal density: \( T(\theta_{here}, \theta_{there}) \) (e.g., move one-at-time random walk)

- **Input:** \( \Lambda, \theta_{old} \) (and \( X \))
- **Propose:** Draw \( \theta_{prop} \) from \( T(\theta_{old}, \theta) \)
- **Accept-Reject:** Put \( \theta^{(i)} \) to \( \theta_{prop} \) 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 \( \theta_{old} \).
- **Repeat:** If \( i < K \) put \( \theta_{old} = \theta^{(i)} \) and go to Propose;
- **else** output \( \theta^{(K)} \).
Next:
Why Does this Work?

- Prove that the particle filter works using the notion of Gallant, A. Ronald, and Han Hong (2007), “A Statistical Inquiry into the Plausibility of Recursive Utility,” *Journal of Financial Econometrics*, that GMM induces a probability space; – next several slides


Joint Density Induced by GMM, Dice Example

Table 4. Tossing two correlated dice \((X, \Lambda)\) when the probability of the difference \(D = X - \Lambda\) is the primitive.

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

Conditional probability is \(P(D = d | \Lambda = \lambda) = P(C_d \cap O_\lambda)/P(O_\lambda)\), where \(O_\lambda\) is the union of the events that can occur. \(Q(\Lambda = \lambda) = P(O_\lambda)\) is the marginal in the sense that \(P(D = d) = \sum_{\lambda=1}^{6} P(D = d | \Lambda = \lambda)Q(\Lambda = \lambda)\)
Conditional Density, Dice Example, 1 of 2

- Let $\mathcal{C}$ be the smallest $\sigma$-algebra that contains the preimages in Table 4.
- Any $\mathcal{C}$-measurable $f$ must be constant on the preimages.
- For such $f$ the formula

$$E(f \mid \Lambda = 2) = \sum_{x=1}^{6} f(x, 2) \sum_{d=-5}^{5} I_{C_d}(x, 2) P(D = d \mid \Lambda = 2)$$

(5)

can be used to compute conditional expectation because $f$ can be regarded as a function of $d$ and the right hand side of (5) equals

$$\sum_{d=-5}^{5} f(d) P(D = d \mid \Lambda = 2).$$
Conditional Density, Dice Example, 2 of 2

- Equation (5) implies that we can view $P(D = d)$ as defining a conditional density function

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

(6)

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

- To get an expression that agrees with the expressions in Gallant and Hong (2007) note that we can write equation (6) as

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

(7)

- Similarly,

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

(8)
Abstraction

• A GMM criterion $Z(X, \Lambda, \theta)$ defines a probability space
  \[(\mathbb{R}^{\text{dim}(X)} \times \mathbb{R}^{\text{dim}(\Lambda)}, \mathcal{C}, P_\theta)\]
  
  – $\mathcal{C}$ is the smallest $\sigma$-algebra containing the preimages of $Z$

• On which there are notions of joint
  
  \[p(X, \Lambda, \theta) = (2\pi)^{-M/2} \exp\left\{ -\frac{1}{2} g_T(X, \Lambda, \theta)' [\Sigma(X, \Lambda, \theta)]^{-1} g_T(X, \Lambda, \theta) \right\},\]
  conditional $p(X | \Lambda, \theta)$, and marginal densities.

• If $P_\theta^o$ denotes the data generating process, then
  \[(\mathbb{R}^{\text{dim}(X)} \times \mathbb{R}^{\text{dim}(\Lambda)}, \mathcal{C}, P_\theta) = (\mathbb{R}^{\text{dim}(X)} \times \mathbb{R}^{\text{dim}(\Lambda)}, \mathcal{C}, P_\theta^o)\]
What if One Knows a Marginal?, Dice Example

• Then one knows the probabilities $P(R_\lambda)$ of the rectangles

\[
R_\lambda = \mathbb{D} \times \{\lambda\} \\
\mathbb{D} = \{1, 2, 3, 4, 5, 6\}
\]

• Let $\mathcal{C}^*$ be the smallest $\sigma$-algebra containing $\{C_d\}_{d=-5}^5$ and $\{R_\lambda\}_{\lambda=1}^6$

• The singleton sets $\{(x, \lambda)\}$ are in $\mathcal{C}^*$ so joint probability $P^*$ on $\mathcal{C}^*$ and conditional densities have their conventional definition

\[
- P^*(X = x \mid \Lambda = \lambda) = \frac{P^*(\{(x, \lambda)\})}{P^*(R_\lambda)} \\
- P^*(\Lambda = \lambda \mid X = x) = \frac{P^*(\{(x, \lambda)\})}{P^*(R_x)}
\]
Indeterminacy, Dice Example, 1 of 2

For $P^*\{(x, \lambda)\}$ we have nine equations in sixteen unknowns:

\[
\frac{4}{18} = \sum_{i=1}^{5} P^*\{(i, i + 1)\}
\]

\[
\frac{10}{18} = \sum_{i=1}^{6} P^*\{(i, i)\}
\]

\[
\frac{4}{18} = \sum_{i=1}^{5} P^*\{((i + 1, i))\}
\]

\[
\frac{1}{6} = P^*\{((1, 1))\} + P^*\{((2, 1))\}
\]

\[
\frac{1}{6} = P^*\{((1, 2))\} + P^*\{((2, 2))\} + P^*\{((3, 2))\}
\]

\[
\frac{1}{6} = P^*\{((2, 3))\} + P^*\{((3, 3))\} + P^*\{((4, 3))\}
\]

\[
\frac{1}{6} = P^*\{((3, 4))\} + P^*\{((4, 4))\} + P^*\{((5, 4))\}
\]

\[
\frac{1}{6} = P^*\{((4, 5))\} + P^*\{((5, 5))\} + P^*\{((6, 5))\}
\]

\[
\frac{1}{6} = P^*\{((5, 6))\} + P^*\{((6, 6))\}
\]

There is one linear dependency leaving eight equations in sixteen unknowns.
Indeterminacy, Dice Example, 2 of 2

• The fact that for $P^*(\{(x, \lambda)\})$ we have only eight equations in sixteen unknowns is fatal.

• We have no logical basis for choosing a particular solution.

• The particle filter depends on the choice of solution.
A Second Example, Mimics Fisher (1930), 1 of 2

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

- The preimages of \( Z(x, \lambda) \) are
  \[ C_z = \{(x, \lambda) : x = z + \lambda, \lambda \in \mathbb{N}\} \quad z \in \mathbb{N} \]
  which lie on 45 degree lines in the \((x, \lambda)\) plane.

- Given \( \lambda \), for every \( z \in \mathbb{N} \) there is an \( x \in \mathbb{N} \) with \((x, \lambda) \in C_z\) so every \( C_z \) can occur. Therefore \( O_\lambda = \cup_{z \in \mathbb{N}} C_z \) and \( P(O_\lambda) = 1 \) for every \( \lambda \in \mathbb{N} \).
If $P(O_{\lambda}) = 1$ for every $\lambda \in \mathbb{N}$.

Then

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

which does not depend on $\lambda$.

Consequently,

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

Provides a rationale for choosing a solution: The conditional probability of $X$ given $\Lambda$ should be the same under $P^*_\theta$ and $P_\theta$.

$$P^*(X = x, \Lambda = \lambda) = P(Z = x - \lambda) P^*(R_\lambda)$$

$$P^*(X = x \mid \Lambda = \lambda) = P(Z = x - \lambda).$$
Abstraction, 1 of 3

- A GMM criterion $Z(X, \Lambda, \theta)$

- And knowledge of $P^o(R_B)$
  - $R_B = \mathbb{R}^{\dim(X)} \times B$
  - $B \in \mathbb{R}^{\dim(\Lambda)}$ is Borel

- Defines a probability space
  $$(\mathbb{R}^{\dim(X)} \times \mathbb{R}^{\dim(\Lambda)}, C^*, P^\theta)$$
  - $C^*$ is the smallest $\sigma$-algebra containing the preimages of $Z$ and rectangles $R_B$,

- On which there are notions of joint $p^*(X, \Lambda, \theta)$, conditional $p^*(X \mid \Lambda, \theta)$, and marginal densities $p^*(\Lambda)$. 
Abstraction, 2 of 3

- \((\mathbb{R}^{\dim(X)} \times \mathbb{R}^{\dim(\Lambda)}, C, P_\theta) = (\mathbb{R}^{\dim(X)} \times \mathbb{R}^{\dim(\Lambda)}, C, P_\theta^o)\)

- \((\mathbb{R}^{\dim(X)} \times \mathbb{R}^{\dim(\Lambda)}, C, P_\theta^*) = (\mathbb{R}^{\dim(X)} \times \mathbb{R}^{\dim(\Lambda)}, C, P_\theta^o)\)

- \((\mathbb{R}^{\dim(X)} \times \mathbb{R}^{\dim(\Lambda)}, C^*, P_\theta^*) = (\mathbb{R}^{\dim(X)} \times \mathbb{R}^{\dim(\Lambda)}, C^*, P_\theta^o)\)
Abstraction, 3 of 3

- If we assume that the union $O_{\lambda}$ of all sets in $\mathcal{C}$ that can occur if $\Lambda = \lambda$ is known to have occurred has probability one, then

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

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

- And we can recover

$$\int f(x, \lambda) p(\lambda | X, \theta) \, d\lambda$$

via a particle filter as long as we restrict attention to $\mathcal{C}$-measurable $f$. 
Interpretation

• If we assume compact $\Theta$, then

• Chernozukov-Hong are Bayes on

$$\left(\mathbb{R}^{\text{dim}(X)} \times \mathbb{R}^{\text{dim}(\Lambda)}, C, P_\theta\right) = \left(\mathbb{R}^{\text{dim}(X)} \times \mathbb{R}^{\text{dim}(\Lambda)}, C, P_\theta^o\right)$$

• And we are Bayes on

$$\left(\mathbb{R}^{\text{dim}(X)} \times \mathbb{R}^{\text{dim}(\Lambda)}, C^*, P_\theta^*\right) = \left(\mathbb{R}^{\text{dim}(X)} \times \mathbb{R}^{\text{dim}(\Lambda)}, C^*, P_\theta^o\right)$$
Contribution

• The contribution of GMM (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.