Simulation-Based Estimation of Continuous Time Models in R

R/Finance 2010

Eric Zivot
University of Washington

Joint work with:

Peter Fuleky
University of Hawaii
Introduction

Goal: Estimate parameters of continuous time diffusion model from discretely sampled data

\[ dy_t = \alpha(y_t, \theta)dt + \sigma(y_t, \theta)dW_t, \quad dW_t \sim \text{iid } N(0, dt) \]

Examples

OU: \[ dy_t = (\theta_0 - \theta_1 y_t)dt + \theta_2 dW_t, \quad \alpha(y_t, \theta) = \theta_0 - \theta_1 y_t, \quad \sigma(y_t, \theta) = \theta_2 \]

CIR: \[ dy_t = (\theta_0 - \theta_1 y_t)dt + \theta_2 \sqrt{y_t} dW_t, \quad \alpha(y_t, \theta) = \theta_0 - \theta_1 y_t, \quad \sigma(y_t, \theta) = \theta_2 \sqrt{y_t} \]

GCIR: \[ dy_t = (\theta_0 - \theta_1 y_t)dt + \theta_2 y_t^{\gamma} dW_t, \quad \alpha(y_t, \theta) = \theta_0 - \theta_1 y_t, \quad \sigma(y_t, \theta) = \theta_2 y_t^{\gamma} \]
Estimation Methods

- MLE – often not feasible
- MLE of approximated model – difficult
- QMLE of discretized model – easy but biased
- GMM – inefficient and biased
- Bayesian MCMC Methods - promising
- Indirect Inference – Corrects bias in QMLE – focus of talk
Indirect Inference

• Distance-based methodology (aka II) developed by Smith (1993), Gourieroux, Monfort, and Renault (1993)
• Score-based methodology (aka EMM) developed by Gallant and Tauchen (1996)
• Computationally less intensive
  (Gallant and Tauchen, 1996; Chumacero, 2001)
II

- Smaller bias and MSE in MA models (Ghysels, Khalaf and Vodounou, 2003)

EMM

- Computationally less intensive (Gallant and Tauchen, 1996; Chumacero, 2001)
- Smaller bias and MSE in MA models (Ghysels, Khalaf and Vodounou, 2003)
- More accurate inference for AR models (Duffee and Stanton, 2008)
- Computationally less intensive (Gallant and Tauchen, 1996; Chumacero, 2001)
Research Agenda and R Contribution

- Implement indirect inference estimation techniques for some commonly used continuous time models (e.g., OU, CIR, etc.)
- Provide systematic comparison and evaluation of different estimators
- Create indirectInference R package
- Give practical advice on use of techniques
Indirect Inference Set-up

\( \{ y_t \}_{t=\Delta}^{n\Delta} \) observations with observation interval \( \Delta \)

Structural model: \( F_{\theta}, \; \theta \in \mathbb{R}^p \), stationary and ergodic

Auxiliary model: \( \tilde{F}_{\mu}, \; \mu \in \mathbb{R}^r, \; r \geq p \)

\( \tilde{\mu} = \arg \max_{\mu} Q_n \left( \{ y_t \}_{t=\Delta}^{n\Delta}, \mu \right) \), where

\[
Q_n = \frac{1}{n-m} \sum_{t=(m+1)\Delta}^{n\Delta} \tilde{f}(y_t; x_{t-\Delta}, \mu), \quad x_{t-\Delta} = \{ y_i \}_{i=t-m\Delta}^{t-\Delta}
\]

\( \tilde{f}(y_t; x_{t-\Delta}, \mu) = \) conditional log density of \( y_t \) for the model \( \tilde{F}_{\mu} \)

\( \mu(\theta) = \arg \max_{\mu} E_{F_{\theta}}[\tilde{f}(y_t; x_{t-\Delta}, \mu)] = \lim p \tilde{\mu} \) under \( F_{\theta} \)

= binding function
Example: OU Model

\[ F_\theta : d y_t = (\theta_0 - \theta_1) dt + \theta_2 dW_t, \quad \theta_i > 0, \quad p = 3, \quad \Delta = 1/52 \]

\[ y_t = \frac{\theta_0}{\theta_1} \left(1 - e^{-\theta_1 \Delta}\right) + e^{-\theta_1 \Delta} y_{t-\Delta} + \theta_2 \sqrt{\frac{1 - e^{-2\theta_1 \Delta}}{2\theta_1}} z_t, \quad z_t \sim \text{iid } N(0,1) \]

\[ \tilde{F}_\mu : y_t = y_{t-\Delta} + (\mu_0 + \mu_1 y_{t-\Delta}) \Delta + \mu_2 \sqrt{\Delta} \epsilon_{t-\Delta} \]

\[ = \mu_0 \Delta + (1 - \mu_2 \Delta) y_{t-\Delta} + \mu_2 \sqrt{\Delta} \epsilon_{t-\Delta}, \quad \epsilon_{t-\Delta} \sim \text{iid } N(0,1), \quad r = 3 \]

\[ \mu(\theta) = p \lim \tilde{\mu} \neq \theta \]

\[ \mu_0(\theta) = \frac{\theta_0}{\theta_1 \Delta} \left(1 - e^{-\theta_1 \Delta}\right), \quad \mu_1(\theta) = \frac{1}{\Delta} \left(1 - e^{-\theta_1 \Delta}\right), \quad \mu_2(\theta) = \theta_2 \sqrt{\frac{1 - e^{-2\theta_1 \Delta}}{2\theta_1 \Delta}} \]
Example: OU Model

• Estimating the “crude Euler” auxiliary model leads to biased estimates (Lo, 1988)
  - Asymptotic discretization bias = $\mu(\theta) - \theta$
  - $\mu(\theta) - \theta \to 0$ as $\Delta \to 0$

• $\mu(\theta)$ is invertible giving analytic II estimates

\[
\hat{\theta}^{II} = \mu^{-1}(\tilde{\mu})
\]
\[
\hat{\theta}_0^{II} = \frac{\tilde{\mu}_0}{\tilde{\mu}_1 \Delta} \ln(1 - \tilde{\mu}_1 \Delta), \quad \hat{\theta}_1^{II} = \frac{-1}{\Delta} \ln(1 - \tilde{\mu}_1 \Delta),
\]
\[
\hat{\theta}_2^{II} = \tilde{\mu}_2 \sqrt{\frac{2 \ln(1 - \tilde{\mu}_1 \Delta)}{1 - e^{\ln(1 - \tilde{\mu}_1 \Delta)}}}
\]
Non-simulation based Estimation

- Assume $\mu(\theta)$ is known (very rare!)
- EMM is GMM with population moment
  
  $$E_{F_\theta} \left[ \frac{\partial \tilde{f}(y_i; x_{i-\Delta}, \mu)}{\partial \mu} \right]_{\mu=\mu(\theta)} = 0$$

- II minimizes distance between $\mu(\theta)$ and $\tilde{\mu}$
- Asymptotically equivalent to MLE when auxiliary model encompasses structural model
Simulation-based EMM and II

- $\mu(\theta)$ is unknown
- $\tilde{\mu}$ is used to estimate $\mu(\theta_{\text{true}})$
- With EMM, simulations for a given $\theta$ are used to approximate the expectation of sample score
- With II, simulations are used to approximate $\mu(\theta)$ for any $\theta$
- Gouriéroux and Monfort (1996) describe 3 types of II estimators and 2 types of EMM estimators
1\textsuperscript{st} Type of Simulation-based II: IL

Real data: \( \{y_t\}_{t=1,\ldots,n} \)

Simulated data: \( \{y_t(\theta)\}_{t=1,\ldots,S_n} \)

Simulations of pseudo-data from the model \( F_\theta \)

Computation of the auxiliary estimator of the model \( \tilde{F}_\mu \)

Distance \( \tilde{\mu} \)

Distance \( \tilde{\mu}_S^L(\theta) = \arg \max_\mu \tilde{Q}_{S_n}(\{y_t(\theta)\}, \mu) \)

\( \hat{\theta}_{S}^{IL}(\Omega) = \arg \min_\theta (\tilde{\mu} - \tilde{\mu}_S^L(\theta))^\prime \Omega (\tilde{\mu} - \tilde{\mu}_S^L(\theta)) \)
2$^{nd}$ Type of II: IM

Real data: $\{y_t\}_{t=1,...,n}$

Simulated data: $\{y^s_t(\theta)\}_{t=1,...,n}^{s=1,...,S}$

Simulations of pseudo-data from the model $F_\theta$

Computation of the auxiliary estimator of the model $\tilde{F}_\mu$

Distance:

$\tilde{\mu} \rightarrow \tilde{\mu}^M_S(\theta) = \frac{1}{S}\sum_s \arg \max_\mu \tilde{Q}_n(\{y^s_t(\theta)\}, \mu)$

$\hat{\theta}^M_S(\Omega) = \arg \min_\theta (\tilde{\mu} - \tilde{\mu}^M_S(\theta))^\prime \Omega (\tilde{\mu} - \tilde{\mu}^M_S(\theta)) \rightarrow \theta$
3\textsuperscript{rd} Type of II: IA

Real data: \( \{y_t\}_{t=1,\ldots,n} \)

Simulated data: \( \{y^s_t(\theta)\}_{t=1,\ldots,n} \)

Simulations of pseudo-data from the model \( F_\theta \)

Computation of the auxiliary estimator of the model \( \widetilde{F}_\mu \)

\[ \widehat{\mu}_{SA}^A(\theta) = \arg \max_\mu \frac{1}{S} \sum_s \bar{Q}_n (\{y^s_t(\theta)\}, \mu) \]

Distance

\[ \hat{\phi}_{SA}^A (\Omega) \equiv \arg \min_\theta (\widehat{\mu} - \widehat{\mu}_{SA}^A(\theta))' \widetilde{\Omega} (\widehat{\mu} - \widehat{\mu}_{SA}^A(\theta)) \]
1\textsuperscript{st} Type of EMM: EL

Real data: $\{y_t\}_{t=1,\ldots,n}$

Simulated data: $\{y_t(\theta)\}_{t=1,\ldots,S_n}$

Simulations of pseudo-data from the model $\mathbf{F}_\theta$

Computation of the auxiliary estimator and score of the model $\mathbf{F}_\mu$

Distance

$\tilde{g}_{Sn}(\theta, \tilde{\mu}) = \frac{\partial \hat{Q}_{Sn}}{\partial \mu} (\{y_t(\theta)\}_{t=1,\ldots,S_n, \tilde{\mu}})$

$\ldots \hat{\theta}_{EL}^E (\Sigma) = \arg \min_{\theta} (\tilde{g}_{Sn}(\theta, \tilde{\mu}))' \Sigma (\tilde{g}_{Sn}(\theta, \tilde{\mu})) \ldots \theta$
2nd Type of EMM: EA

Real data: \( \{ y_t \}_{t=1,...,n} \)

Simulated data: \( \{ y_t^s(\theta) \}_{t=1,...,n} \)

Simulations of pseudo-data from the model \( F_\theta \)

Computation of the auxiliary estimator and score of the model \( \hat{F}_\mu \)

Distance

0

\[ g_n^s(\theta, \bar{\mu}) = \frac{\partial \hat{Q}_n}{\partial \mu}(\{y_t^s(\theta)\}_{t=1,...,n, \bar{\mu}}) \]

\[ \hat{\theta}_S^{EA}(\bar{\Sigma}) = \arg \min_\theta \left( \frac{1}{S} \sum_{s=1}^{S} g_n^s(\theta, \bar{\mu}) \right)' \beta \left( \frac{1}{S} \sum_{s=1}^{S} g_n^s(\theta, \bar{\mu}) \right) \]
R Implementation of II

- Estimate Euler auxiliary model parameters $\mu$ from observed data $\{y_t\}$ by QMLE

$$y_{t+\Delta} - y_t = \alpha(y_t, \mu) \Delta + \sigma(y_t, \mu) \sqrt{\Delta} z_t, \quad z_t \sim \text{iid } N(0,1)$$

$$\tilde{\mu} = \arg\max_{\mu} Q_n \left( \{y_t\}_{t=\Delta}^{n\Delta}, \mu \right), \text{ where}$$

$$\tilde{Q}_n = \frac{1}{n - m} \sum_{t=(m+1)\Delta}^{n\Delta} \tilde{f}(y_t; x_{t-\Delta}, \mu), \quad x_{t-\Delta} = \{y_i\}_{i=t-m\Delta}^{t-\Delta}$$

- Use function `EULEReloglik()` from R package sde

- Use R function `optim()`
R Implementation of II

• Simulate from $F_\theta$
  – In general, cannot do exact simulations because transition density is not known
  – Simulate from very fine Euler discretization
  – Use function `sde.sim()` from R package `sde`
  – Use custom C code for fast simulation
  – Need to worry about “inadmissible” or “explosive” simulations from inappropriate $\theta$ – need to “bullet proof” the simulator
R Implementation of II

• For distance-based II, estimate binding function $\mu(\theta)$ from simulated data $\{y_t^s(\theta)\}$

$$\tilde{\mu}^L_S = \arg\max_{\mu} Q_{Sn} \left( \{y_t^s(\theta)\}, \mu \right), \text{ where}$$

$$\tilde{Q}_{Sn} = \frac{1}{n'} \sum_{t=(m+1)\Delta}^{Sn\Delta} \tilde{f}(y_t^s(\theta); x_{t-\Delta}^s(\theta), \mu),$$

– Use same random number seed for all $\theta$
R Implementation of II

- For score-based II, estimate auxiliary score from simulated data \( \{y_t^s(\theta)\} \) and evaluate at auxiliary parameter estimate

\[
g_{sn}(\{y_t^s(\theta)\}, \tilde{\mu}) = \frac{\partial Q_{sn}(\{y_t^s(\theta)\}, \tilde{\mu})}{\partial \mu}
\]

- User specified function to evaluate score function
- Use same random number seed for all \( \theta \)
R Implementation of II

• For distance-based II, estimate $\theta$

\[
\hat{\theta}^D_s = \arg \min_\theta (\bar{\mu} - \bar{\mu}_s(\theta))' \tilde{\Omega}(\bar{\mu} - \bar{\mu}_s(\theta)), \ i = L, M
\]

• For score-based II, estimate

\[
\hat{\theta}^{EMM}_s = \arg \min_\theta g_{sn}(\theta, \bar{\mu})' \tilde{\Sigma} g_{sn}(\theta, \bar{\mu})'
\]

• If $p = r$ then use identity matrix for weight matrix

• For optimization, use R function `optim()` with Nelder-Meade simplex algorithm
Illustration

- OU Process calibrated to US interest rates used by Phillips and Yu (2009)
  \[ \theta_0 = 0.01, \quad \theta_1 = 0.10, \quad \theta_2 = 0.10 \]
  \[ \frac{\theta_0}{\theta_1} = 0.10 = \text{annualized avg rate,} \]
  \[ \theta_1 = 0.1 \Rightarrow 7 \text{ year half of rate shock} \]
  \[ \theta_2 = 0.10 = \text{annualized rate volatility} \]
  \[ T = 19.23, \quad \Delta = 1/52 \Rightarrow n = 1000 \]

- \( \theta_1 \) is the most difficult parameter to estimate
Shape of distance-based II Objective function

© Eric Zivot 2010
Shape of Gallant-Tauchen Score-based II Objective Function

\[ J_{\text{stat}} \]

\[ \theta_1 \]

\[ \theta_0 \]

© Eric Zivot 2010
95% Confidence Intervals for $\theta_1$

Note: $S = 20$ for simulation-based estimates
Impact of simulation Size S on Densities of $\theta_1$ Estimates
Distribution of $\hat{\theta}_1$: $\theta = (0.01, 0.1, 0.1)$, $n=1000$, $\Delta = 1/52$
Distribution of \( \hat{\theta}_0 \): \( \theta = (0.01, 0.1, 0.1) \), \( n=1000 \), \( \Delta = 1/52 \)
Distribution of $\hat{\theta}_2$: $\theta = (0.01, 0.1, 0.1)$, $n=1000$, $\Delta = 1/52$
Rejection frequencies of likelihood ratio type tests at 5% nominal level of significance for 1000 Monte Carlo simulations.

<table>
<thead>
<tr>
<th>EN1</th>
<th>EL1</th>
<th>EA1</th>
<th>EN2</th>
<th>EL2</th>
<th>EA2</th>
<th>EM2</th>
<th>IN</th>
<th>IL</th>
<th>IA</th>
<th>IM</th>
</tr>
</thead>
</table>

True $\theta = (0.01, 0.1, 0.1)$, $n = 1000$, $\Delta = 1/52$.

<table>
<thead>
<tr>
<th>Joint Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.767 0.760 0.728 0.129 0.114 0.116 0.419 0.121 0.112 0.113 0.415</td>
</tr>
</tbody>
</table>

Simple Test of $\theta_1 = 0.1$

| 0.698 0.640 0.593 0.197 0.169 0.175 0.130 0.210 0.176 0.183 0.149 |
Research in Progress

References


References

