Simulation Estimation for Dynamic Latent Variable Models

David Childers

September 27, 2011

Outline

1. Model
   - Description
   - Estimation Methods
   - Restrictions
2. Indirect Inference
   - Description
   - Properties
3. Binding Functions
   - Role
   - Non- and semiparametric Methods
4. Performance Issues
5. Extensions and Conclusions
   - Areas for Improvement
   - Conjectures and Conclusions

The Model

Dynamic Latent Variable Model

$\tilde{Y}_{t+1} = r(\tilde{Y}^t, \varepsilon_t, \theta)$

$\theta \in \Theta \in \mathbb{R}^q$, $\widetilde{Y}_t = (y_t', w_t')' \in \mathbb{R}^k \times \mathbb{R}^{p-k}, \tilde{Y}^t = (\widetilde{Y}_t, \tilde{Y}_{t-1}, ..., \tilde{Y}_1)$

ε_t white noise

$\{y_t\}_{t=1}^T$ observed, $\{w_t\}_{t=1}^T$ not observed: “Latent”

Names for this model: “State Space Model,” “Dynamic Latent Variable Model,” “Hidden Markov Model”

Many variants, depending on functional form, timing, variables included, etc

Why this model?

- Generality: large class of time series models
- Theoretical reasons
  - Economic model has role for y_t and w_t, data contains only y_t
  - Why not just use reduced form?
    - Fine for forecasting
    - Parameter inference?
    - Counterfactuals?
  - Linear case: VAR vs. linear state space model
    - If system stochastically nonsingular and invertible, have $VAR(\infty)$ representation
    - Causal interpretation for coefficients?

Why this model?

- Practical Reasons
  - Empirically, finite order VARs for macro data have non-white noise errors
  - Implementation issues for infinite order VARs
  - Data may have features not captured in Wold representation: higher order moments, conditional moments, etc

Direct Estimation

- Exact likelihood available in many cases, for MLE or Bayes
  - No hidden state: standard Markov models
  - Linear with Gaussian errors: Kalman filter
  - Certain types of ARCH, GARCH, Stochastic Volatility, Diffusion models, linear factor models
- Closed form likelihood often not available, though it can be calculated given conditional likelihoods (here for state space form)
  - $f(y^T|y_0, w_0, \theta) = \int \Pi_{t=1}^T f(y_t|y^{t-1}, w^{t-1}, \theta) f(w_t|y^{t-1}, w^{t-1}, \theta) \Pi_{t=1}^T d\mu(w_t)$
  - Problem: $T * (p-k)$ dimensional integral not always easy to calculate, without closed form or very fast numerical methods

Recursive Formulation

- Some simplification may be afforded by using recursive updating: suppose $f(w_{t-s}^{t-1}|y^{t-1})$ known

Kitagawa's filtering algorithm

1. Obtain $f(w_{t-s}^{t}|y^{t-1}) = f(w_t|y^{t-1}, w_{t-s}^{t-1})f(w_{t-s}^{t-1}|y^{t-1})$
2. Multiply to get $f(y_t, w_{t-s}^{t}|y^{t-1}) = f(y_t|y^{t-1}, w_{t-s}^{t})f(w_{t-s}^{t}|y^{t-1})$
3. Integrate $f(y_t|y^{t-1}) = \int f(y_t, w_{t-s}^{t}|y^{t-1})d\mu(w_{t-s}^{t})$
4. Update $f(w_{t-s}^{t}|y^{t}) = \frac{f(y_t, w_{t-s}^{t}|y^{t-1})}{f(y_t|y^{t-1})}$
5. (If r depends only on past s-1 lags of w) integrate over w_{t-s} to get $f(w_{t-s+1}^{t}|y^{t-1})$
6. Go back to 1., until $f(y_t|y^{t-1})$ known for all $t = 1, ..., T$

- If the dependence on lagged values is of a relatively small order, step 3 may be relatively low dimensional integration
- Explicit formulas for finite state HMMs, Markov switching models
- Still requires integrating T times: if done numerically, error compounds over time, approximation becomes arbitrarily bad

Approximation Methods

- In general case, exact likelihood methods not feasible: consider approximations
- What to approximate:
  - The model itself
  - The likelihood function

Approximating the Model

- Observe the hidden state: use some noisy measurement and then estimate standard model
    - Adds measurement error, but often easy to resolve with standard techniques
    - Popular for stochastic volatility models, some macro variables
- Discretize into finite points
    - Likelihood now exists, but need number of points small relative to number of observations for estimation
- Perturbation: $\widetilde{Y}_{t+1} = r(\tilde{Y}_o) + \frac{dr}{d\tilde{Y}}(\tilde{Y}_o)\tilde{Y}_t + \frac{1}{2}\tilde{Y}_t' \frac{d^2 r(\tilde{Y}_0)}{d\tilde{Y}^2}\tilde{Y}_t + ...$
    - Extremely popular method: first order approximation (linearization or log linearization) has exact likelihood (extended Kalman filter)
    - Beyond first or second order, no longer have likelihood function

Approximating the Likelihood

- Numerical integration
  - Quadrature: slow
  - MCMC: not quite as slow, in some cases
- EM algorithm
- Simulated maximum likelihood
- All of the above are fine, but require the approximation to be arbitrarily close (relative to sampling error) to yield consistent estimates
  - Specifically, for SMLE, consistency requires $\frac{S}{T} \longrightarrow \infty$ and asymptotic normality requires $\frac{\sqrt{S}}{T} \longrightarrow \infty$

General Approximation Estimators

- Kristensen and Salanié (2009)
- Suppose first derivative of criterion function for consistent, asymptotically normal estimator is given by $G(y^T, \theta_0, \gamma_0) = o_p(\frac{1}{\sqrt{T}})$ with γ possibly infinite dimensional
- Estimator with no approximation satisfies $\hat{\theta}_T - \theta_0 = -\frac{dG}{d\theta}^{-1} G(y^T, \theta_0, \gamma_0) + o_p(\frac{1}{\sqrt{T}})$
- Consider approximation of model or criterion satisfying $\hat{\gamma}_T \to \gamma_0$, so $\tilde{\theta}_T - \theta_0 = -\frac{dG}{d\theta}^{-1} G(y^T, \theta_0, \hat{\gamma}_T) + o_p(\frac{1}{\sqrt{T}})$
- Then $\tilde{\theta}_T - \hat{\theta}_T \simeq -\frac{dG}{d\theta}^{-1}(G(y^T, \theta_0, \hat{\gamma}_T) - G(y^T, \theta_0, \gamma_0)) \simeq -\frac{dG}{d\theta}^{-1}(\nabla G(y^T, \theta_0, \gamma_0)[d\hat{\gamma}] + \nabla^2 G(y^T, \theta_0, \gamma_0)[d\hat{\gamma}, d\hat{\gamma}] + remainder)$
- Mathematically identical to two-step sieve estimation

Disadvantages of Approximation Methods

- May not even be able to find approximator of likelihood function in closed form
- Except in very special cases (linearity), bias not asymptotically negligible unless approximation arbitrarily close
- Often, one can simulate draws from a model without knowing the likelihood, and even this may be computationally burdensome to do repeatedly
- Want methods that work reasonably well with intractable functions and moderate computational difficulty

Why not tractable?

Dynamic economic models present particular difficulties beyond unobserved states which make repeated simulation burdensome

- Rational expectations
  - transition function includes expectation with respect to the true model
- Optimization
  - Policy function generally arg max of complicated function
- Equilibrium:
  - May generate a fixed point problem

Computational methods exist, also econometric methods
Want to economize on knowledge of functional forms and number of simulations

Partial Solutions

- If functional form of likelihood unknown, can use NPSMLE (Fermanian & Salanié 2004, Kristensen and Shin 2008)
    - Simulate model large number of times per parameter value, estimate likelihood by kernel density, maximize estimated likelihood
    - Requires fast rate of simulation growth for consistency and normality
- Flury and Shephard (2008) show particle filter with Metropolis-Hastings algorithm can generate draws from exact likelihood if conditional likelihood of observation equation is known, using only simulations from state transition equation
    - Noise added by particle filter adds variance, does not induce bias
    - Allows calculation of posterior for Bayesian estimation by MCMC
    - Posterior mean is consistent estimator by Bernstein-Von Mises Theorem
    - As a result, mean of MCMC draws is a consistent estimator without taking simulations to infinity
    - Gets around problems with filtering algorithm when used in MLE
- Above method probably ideal if observation equation has known likelihood

Indirect Inference

Basic Idea

- Find some function of the data $\hat{\beta}_T$, generally some kind of extremum estimator
- Simulate data sets from model at a given parameter value
- Estimate same function on simulated data, $\hat{\beta}_s(\theta)$, S times
- Find θ which minimizes distance between $\hat{\beta}_T$ and averaged $\hat{\beta}_s(\theta)$
- Under weak regularity conditions, obtain consistent asymptotically normal estimator of θ_0 with fixed number of simulations
- History:
  - Smith (1990, 1993), Gouriéroux, Monfort, Renault (1993)
  - Special case: SMM: McFadden (1989), Pakes and Pollard (1989)

The Estimator

Indirect Inference Estimator

$$\hat{\theta}_T^{II} \equiv arg\ \inf_{\theta\in\Theta}[(\hat{\beta}_T - \tfrac{1}{S}\textstyle\sum_{s=1}^{S} \hat{\beta}_s(\theta))'\hat{\Omega}_T(\hat{\beta}_T - \tfrac{1}{S}\textstyle\sum_{s=1}^{S} \hat{\beta}_s(\theta))] + o_p(\tfrac{1}{d_T})$$

- Simulated estimator is averaged over S simulations
- $\hat{\Omega}_T$ is some pos. def. weight matrix converging in probability to Ω
- As standard for extremum estimation, approximation is allowed

How does it work?

- High level conditions
  - $\hat{\beta}_T \xrightarrow{p} \beta(\theta_o)$ and $\hat{\beta}_s(\theta) \xrightarrow{p} \beta(\theta)$ uniformly over compact parameter space Θ, where $\beta(\theta)$ is a continuous, differentiable, one-to-one function of θ
    - needs to convey information about θ
  - $d_T(\hat{\beta}_T - \beta(\theta_0)) \xrightarrow{d} N(0, I_0)$ and $d_T(\hat{\beta}_s(\theta_0) - \beta(\theta_0)) \xrightarrow{d} N(0, I_0)$ for some asymptotic covariance matrix I_0 and rate $d_T \to \infty$
    - Usually these are extremum estimators, with sandwich covariance matrix, or M-estimators, and so $d_T = \sqrt{T}$
  - Although simulations are generally drawn independent of data, may allow simulated and observed estimators to be correlated, e.g., due to conditioning on exogenous variables, in which case $Cov(d_T\hat{\beta}_T, d_T\hat{\beta}_s(\theta_0)) \xrightarrow{p} K_0$
  - $\hat{\beta}_s(\theta)$ is twice differentiable with first derivative at θ_0 converging uniformly in probability to some full rank matrix J and second derivative $O_p(1)$

Consistency

- Consistency follows by uniform convergence in probability to asymptotic criterion function $(\beta(\theta_0) - \beta(\theta))'\Omega(\beta(\theta_0) - \beta(\theta))$
- Continuity, compactness, and one-to-one property of binding function guarantees identification condition
- Uniform convergence in probability generally result of standard assumptions guaranteeing a uniform LLN (for M-estimators) or stochastic equicontinuity in general (for optimization estimators): usually a Lipschitz condition suffices, though for non-smooth criteria may need something like VC subgraphs, bracketing, or fat-shattering conditions)

Asymptotic normality

Taylor expansion around θ_0 gives

$$o_p(1) = -d_T\left(\tfrac{1}{S}\textstyle\sum_{s=1}^{S} \frac{d\hat{\beta}_s(\theta_0)}{d\theta}\right)'\hat{\Omega}_T\left(\hat{\beta}_T - \tfrac{1}{S}\textstyle\sum_{s=1}^{S} \hat{\beta}_s(\theta_0)\right) +$$
$$d_T\left(\left(\tfrac{1}{S}\textstyle\sum_{s=1}^{S} \frac{d^2\hat{\beta}_s(\theta_0)}{d\theta^2}\right)'\hat{\Omega}_T\left(\hat{\beta}_T - \tfrac{1}{S}\textstyle\sum_{s=1}^{S} \hat{\beta}_s(\theta_0)\right) +\right.$$
$$\left.\left(\tfrac{1}{S}\textstyle\sum_{s=1}^{S} \frac{d\hat{\beta}_s(\theta_0)}{d\theta}\right)'\hat{\Omega}_T\left(\tfrac{1}{S}\textstyle\sum_{s=1}^{S} \frac{d\hat{\beta}_s(\theta_0)}{d\theta}\right)\right)(\hat{\theta}_T^{II} - \theta_0)$$

Rearrange and simplify to get

$$d_T(\hat{\theta}_T^{II} - \theta_0) = (J'\Omega J + o_p(1))^{-1} J'\Omega d_T(\hat{\beta}_T - \tfrac{1}{S}\textstyle\sum_{s=1}^{S} \hat{\beta}_s(\theta_0)) + o_p(1)$$

Asymptotic normality of first step estimators then gives

$$d_T(\hat{\theta}_T^{II} - \theta_0) \xrightarrow{d} N(0, (1 + \tfrac{1}{S})(J'\Omega J)^{-1} J'\Omega(I_0 - K_0)\Omega J(J'\Omega J)^{-1})$$

Optimal weighting matrix given by $\Omega^* = (I_0 - K_0)^{-1}$

Alternative Formulations

- Under stationarity and ergodicity, first order asymptotically equivalent to use one simulation of length ST
- In case auxiliary model is a differentiable MLE with score $g(y^T, \beta)$, first-order equivalent to solve
    - $\hat{\theta}_T^{EMM} = arg\ \inf_{\theta \in \Theta} g(\tilde{y}^{ST}(\theta), \hat{\beta}_T)' \Sigma g(\tilde{y}^{ST}(\theta), \hat{\beta}_T)$
- May confer computational advantages, though higher-order properties not known to hold

Alternative Asymptotics

- Derivation depended crucially on linearity and normality, to be able to find distribution as sum of normals
- In the case that the auxiliary estimator is not itself differentiable but its limit is, Pakes and Pollard (1989) demonstrate different limit theory involving local uniform approximation of the Taylor expansion of the criterion function
- More generally, II estimator is given as an (implicit) function of an auxiliary estimator, so can derive asymptotics using the distribution of that first stage estimator and either a delta method or CMT suitably modified to take into account the sample-size dependence of that implicit estimator
  - Permits derivations in case of non-normal or mixed rate asymptotics: e.g. unit roots

Choosing an auxiliary model

- Continuous one-to-one binding function needed for identification, but specific choice up to econometrician
- Goals
  - Parameter identification: prefer flexible form to ensure that variation in θ causes identifiable variations in β
  - Efficiency: prefer form closely matching the data
- Common choice: biased estimator of parameters: ensures identification, also gives close match to true form
  - Ex: MLE of time-discretized or approximated model
  - Allows correction for approximation error
- Or, find moments heuristically thought to be informative

Formal Criteria

Gallant and Tauchen (1996)

- Suppose that the auxiliary model is an MLE with likelihood $p(y^T, \beta)$ and the true model has likelihood $f(y^T, \theta)$

Smooth embedding

If in an open neighborhood of θ_0, there is a twice differentiable mapping $g(\theta)$ from Θ to the auxiliary model parameter space such that $f(y^T, \theta) = p(y^T, g(\theta))$, then the auxiliary model satisfies "smooth embedding"

- If the weight matrix is given by a consistent estimator of $Var(\sqrt{T}\frac{d}{d\beta}p(y^T, \beta(\theta_0))$ and the auxiliary likelihood satisfies smooth embedding, then the "score" form yields an estimate of θ_0 which achieves the Cramer-Rao efficiency bound for the true model
- How to know if this condition is satisfied? Need flexible likelihood
- Gallant and Long (1997): sieve MLE with Hermite polynomials

Conditional moments

- Often difficult to obtain closed form for conditional moments
  - e.g. $E[y_t|y_{t-1}]$ depends in complicated manner on latent variable
- Solution: kernels:
  - Use Nadaraya-Watson regression to estimate conditional moment on simulated and observed data
    - ★ At a finite set of points (Billio & Monfort 2003)
    - ★ At all points, and generate moments by multiplication with instruments (Creel and Kristensen 09)
  - Results technically hold without full kernel convergence if kernel thought of as arbitrary binding function
  - Latter results, given uniform convergence for kernel estimates, can achieve parametric rate if simulations increase with sample size

Distributions

- May have more information in full conditional distributions than just moments
- Altissimo and Mele: Kernel density estimates

$$\arg\ \min_{\theta\in\Theta} \int\int (\frac{1}{S}\sum_{i=1}^{S} \pi_u^i(y|y',\theta) - \pi_u(y|y'))^2 T(y,y',\theta)W(y,y',\theta)dydy'$$

- Achieves parametric rate with S fixed
- Trimming function T needed to handle conditioning on tail events
- Concern: requires numerical integration, though low-dimensional

Model Specification

- Above results assume that simulations draw from the true distribution
- Binding function may not characterize entire distribution of data, but simulation draws on fully specified model
- Want robustness against misspecificaion for “nuisance” features of model
- Can we do semiparametric inference using simulation estimation?

Semiparametric Indirect Inference

Dridi, Guay, and Renault (2007)

- Let data be drawn from infinite dimensional true model parameterized by θ_1^*
- Simulate data from model parameterized by $\theta = (\theta_1', \theta_2')'$, θ_2 set of "nuisance parameters"
- Estimate using auxiliary model yielding binding function $\beta(\theta_1, \theta_2)$

Partial Encompassing

Suppose that for any true probability law there exists a unique set of pseudo-true parameter values θ_2^* such that under the true model, $\hat{\beta} \xrightarrow{p} \beta(\theta_1^*, \theta_2^*)$ uniformly, and this binding function is one-to-one, then the binding function "partially encompasses" the simulated model

- Under partial encompassing, indirect inference provides consistent and asymptotically normal estimates of θ_1.
- Likewise, if a nuisance parameter is fixed at its true value and partial encompassing holds for other parameters, still consistent, asymptotically normal

Interpretation

- Partial encompassing requires that estimates of some parameters not affected by misspecification
- Characterizes robustness, but high level condition
- Most likely to be satisfied for moment conditions in static or linear models
- Nonlinear dynamics
  - Generally affect entire distribution: higher order changes have first and second order effects
  - Problem with any semiparametric inference in this class of models

Efficient Estimation

- Efficient estimation, testing, require optimal weight matrix estimation
- Consistent estimators easy to find: plug-in analogs of asymptotic values
- Usually need HAC/HAR estimates: standard results
- Monte Carlo suggests version with estimated optimal weight matrix has higher MSE than inefficient unweighted model
- Known small sample issues with estimated covariances for dynamic models

Efficient Weight Matrix

- One option: continuous updating version of estimator: entirely analogous to GMM case
- Asymptotically equivalent procedure: indirect likelihood estimator (Creel & Kristensen 2011)
  - Choose asymptotically normal statistic $\hat{\beta}$ with an Edgeworth expansion
  - Maximize log likelihood of $\hat{\beta}$ with respect to θ (or equivalently, calculate posterior mean of θ given $\hat{\beta}$)
  - Since likelihood not known, need to simulate it: need simulations to diverge, as in SML

Small Sample Properties

- Common use of indirect inference is to “bias correct” an approximated auxiliary model
  - Actually a “consistency correction”
- In fact indirect inference can correct small-sample bias
  - As simulations go to ∞, $\frac{1}{S}\sum_{s=1}^{S}\hat{\beta}_T^s \xrightarrow{p} E[\hat{\beta}_T] = b_T(\theta)$, small sample mean
  - Then have $E[b_T(\hat{\theta}_T^{II})] = b_T(\theta_0)$: exact bias correction if $b_T(\theta)$ affine
  - Can iterate if not affine
  - Analogous to median unbiased procedure of Andrews (1993)

Higher Order Properties

- Monte Carlo with models known to be biased, such as MLE for AR(1), finds smaller bias than MLE
- Suppose auxiliary model is consistent and has Edgeworth expansion uniformly in θ
  - $\hat{\beta}_T(\theta) = \theta + \sum_{k=1}^{K} \frac{\pi_k(\theta)}{T^{K/2}} + o_p(T^{-(K+1)/2})$ where $\pi_K(\theta)$ is a random polynomial
- Then equating the Edgeworth expansion of simulated auxiliary model to that of estimated one and taking Taylor expansion around θ_0 gives Edgeworth expansion of $\hat{\theta}_T^{II}$.
- If simulations go to ∞, obtain that expectation of $O_p(\frac{1}{\sqrt{T}})$ and $O_p(\frac{1}{T})$ terms is 0.
- Similar to bootstrap bias correction

Identification

- One-to-one property of binding function not guaranteed for dynamic latent variable models:
    - low order moments generally insufficient to identify arbitrary DLV model
    - Nonparametric methods not necessarily sufficient if only identifying finite period likelihood
- Method can't handle partially identified models
    - Del Negro, Schorfheide, Smets, Wouters (2007) provide similar procedure in Bayesian setting
    - Gallant and McCulloch (2008) update to allow simulation
    - Here, non-identified parts of model simply don't update posterior (except via correlated priors)

Robustness

- Method relies strongly on having complete, parametric model
- Low-level conditions for estimates not to depend on certain aspects of model not worked out
  - Deeper issue of characterizing dependence on distributional assumptions in nonlinear dynamics
- Allowing nonparametric components of model
  - In progress?
- Dependent data? First-step estimators rely on y_t stationary, ergodic, mixing
  - Even if $\tilde{Y}_t$ satisfies this, y_t may not
  - Very few estimators known to work under general conditions: MLE consistent

Approximate Models

- Extremely common to simulate from model which is itself approximated
- Simulating from model approximated by perturbation
- Characterization of dependence on this approximation?
  - Identification conditions characterized for first-order state space approximation
  - Conditions for second, higher orders should be weaker
  - Effect on estimator?

Conjectures I: Objective Function Form

- II estimator takes extremely similar form to GMM
- Many procedures related to GMM easily translated to II setting
  - Testing, instrument selection, Laplace-type estimator equivalence
- What ideas in GMM worth translating to this setting?
  - GEL estimation?
    - Turns out not to be directly feasible: nonlinear term enters inside summation
    - Exception is CUE: quadratic form leads to linear derivatives
  - Dominguez and Lobato estimator?
    - Extremely similar to Altissimo and Mele estimator
    - Just change conditional density estimator to conditional expectation estimator

Conjectures II: Filtering

- Applied researchers often care a lot about the latent variable itself
- Simulation estimators don't naturally provide this, unlike likelihood estimators using filtering
- Is it possible to incorporate a filtering step along with or after estimation step?
- Technically yes: can always do SMLE, or use nonparametric density estimators to find $f(w_t|y^T)$
- Problem is that distribution depends on entire past path: require path simulations to go to infinity
  - Nonparametric conditional estimation on high dimensions has low convergence rate
  - Requiring conditioning on growing dimensions yields even slower rate
  - Much worse than usual SMLE rate, which is already pretty bad
  - Problem doesn't go away with simulation of conditional distributions and then filtering!
  - Errors propagate over time: estimate becomes worse as T grows

Conclusion

- Dynamic latent variable models provide particular challenges to estimation
- Simulation estimation provides straightforward way around these challenges
- Requires some care, but methods exist for consistent and efficient estimates for fairly general models
- With sufficient computational power, method also has good small sample properties