Lectures 10 and 11. Bayesian and Quasi-Bayesian Methods

Fall, 2007

Outline:

1. Informal Review of Main Ideas
2. Monte-Carlo Examples
3. Empirical Examples
4. Formal Theory

References:
Theory and Practice:

Van der Vaart, A Lecture Note on Bayesian Estimation
Chernozhukov and Hong, An MCMC Approach to Classical Estimation, JoE, 2003
Liu, Tian, Wei (2007), JASA, 2007 (forthcoming).

Computing:
Chib, Handbook of Econometrics, Vol 5.
Geweke, Handbook of Econometrics, Vol 5.

Part 1. Informal Introduction An Example (Chernozhukov & Hong, 2003)

Consider GMM estimator for Instrumental Quantile Regression Model:

$$E\big(\tau - 1(Y \leq D'\theta)\big)Z = 0.$$

Maximize criterion

$$L_n(\theta) = -n \underbrace{\frac{1}{2} g_n(\theta)' W(\theta) g_n(\theta)}_{\widehat{Q}(\theta)}$$

with

$$g_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\tau - 1(Y_i \leq D_i'\theta)\big)Z_i$$

and

$$W(\theta) = \frac{1}{\tau(1-\tau)}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i'\right]^{-1}$$

Computing extremum is problematic.

Smoothing does not seem to help much.

Some other examples:

Nonlinear IV & GMM problems with many local optima.

Powell's censored median regression.

Overview of Results:

1. Interpret

$$p_n(\theta) \propto \exp(L_n(\theta))$$

as *posterior* density, summarizing the beliefs about the parameter.

This will encompass the Bayesian learning approach, where $L_n(\theta)$ is proper log-likelihood.

Otherwise treat $L_n(\theta)$ as a "replacement" or "quasi" log-likelihood, and posterior as quasi-posterior.

2. A primary example of an estimator is the posterior mean

$$\widehat{\theta} = \int_\Theta \theta p_n(\theta) d\theta,$$

which is defined by integration, not by optimization. This estimator is asymptotically equivalent to extremum estimator θ^*:

$$\sqrt{n}(\widehat{\theta} - \theta^*) = o_p(1),$$

and therefore is as efficient as θ^* in large samples.

For likelihood framework this was formally shown by Bickel and Yahav (1969) and many others. For GMM and other non-likelihood frameworks, this was formally shown by Chernozhukov and Hong (2003, JoE) and Liu,

Tian, Wei (2007, JASA).

3. When a generalized information equality holds, namely when the Hessian of the objective function $\widehat{Q}(\theta)$ is equal to the variance of the score,

$$\underbrace{\nabla_\theta^2 Q(\theta_0)}_{=:J(\theta_0)} = \underbrace{\text{var}[\sqrt{n}\nabla Q(\theta_0)]}_{=:\Omega(\theta_0)},$$

we can use posterior quantiles of beliefs $p_n(\theta)$ for inference. This is true for the regular likelihood problems and optimally weighted GMM.

4. Numerical integration can be done using Markov Chain Monte Carlo (MCMC), which creates a dependent sample

$$S = (\theta^{(1)}, ..., \theta^{(k)}),$$

a Markov Chain, whose marginal distribution is

$$C \cdot \exp(L_n(\theta)).$$

This is done by using the Metropolis-Hastings or Gibbs algorithms or a combination of the two.

Compute the posterior mean of $S \longrightarrow \widehat{\theta}$. Can also use quantiles of the chain S to form confidence regions.

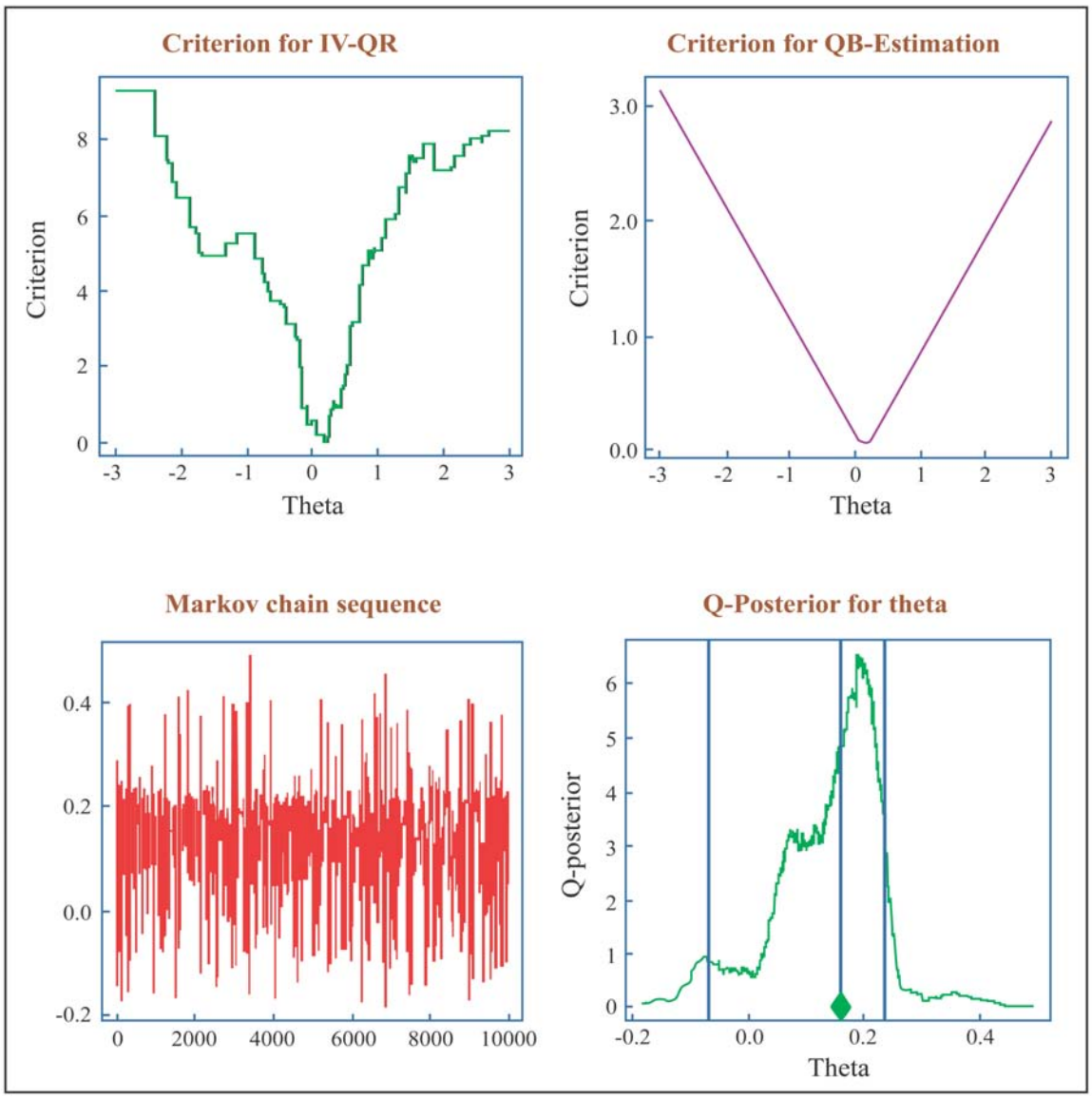


Image courtesy of MIT OpenCourseWare.

Formal Definitions

Sample criterion function

$$L_n(\theta)$$

Motivation of extremum estimators: learning by analogy $\hat{Q} = n^{-1}L_n \to Q$, so extremum estimator $\to \theta_0$, the extremum of Q.

$L_n(\theta)$ is not a log-likelihood function generally, but

$$p_n(\theta) = \frac{\exp[L_n(\theta)]\pi(\theta)}{\int_\Theta \exp[L_n(\theta')]\pi(\theta')d\theta'} \tag{1}$$

or simply

$$p_n(\theta) \propto \exp[L_n(\theta)]\pi(\theta) \tag{2}$$

is a proper density for θ. Treat it as a form of posterior beliefs. Here, $\pi(\theta)$ is a weight or prior density that is strictly positive and continuous over Θ.

Recall that proper posteriors arise from a formal Bayesian learning model:

$$\begin{aligned} p_n(\theta) &= f(\theta|data) = f(data|\theta)\pi(\theta)/f(data) \\ &\propto f(data|\theta)\pi(\theta). \end{aligned}$$

An example of estimator based on posterior is the posterior mean:

$$\widehat{\theta} = \int_\Theta \theta p_n(\theta)d\theta.$$

Definition 1 *The class of QBE minimize the expected loss under the belief* p_n*:*

$$\begin{aligned} \widehat{\theta} &= \arg\min_{d\in\Theta} \left[E_{p_n}(\rho(d-\theta)) \right] \\ &= \arg\min_{d\in\Theta} \left[\int_\Theta \rho(d-\theta) p_n(\theta) d\theta \right], \end{aligned} \tag{3}$$

where $\rho(u)$ is a *penalty or bernoullian loss function*:

i. $\rho(u) = ||u||^2$,

ii. $\rho(u) = \sum_{j=1}^k |u_k|$, an absolute deviation loss,

iii. $\rho(u) = \sum_{j=1}^k (\tau_j - 1(u_j \leq 0))u_j$, loss function.

Loss (i) gives posterior mean as optimal decision.
Loss (ii) gives posterior (componentwise) median as optimal decision.
Loss (iii) gives posterior (componentwise) quantiles as optimal decision.

Computation

Definition 2 ((Random Walk) Metropolis-Hastings)

Given quasi-posterior density $p_n(\theta)$, known up to a constant, generate $\left(\theta^{(0)}, ..., \theta^{(B)}\right)$ by,

1. *Choose a starting value $\theta^{(0)}$.*

2. *For $j = 1, 2, ..., B$, generate $\xi^{(j)} = \theta^{(j)} + \eta^{(j)}$, $\eta^{(j)} \sim N(0, \sigma^2 I)$, and set*

$$\theta^{(j+1)} = \begin{cases} \xi^{(j)} & \text{with probability} \quad \rho(\theta^{(j)}, \xi^{(j)}) \\ \theta^{(j)} & \text{with probability} \quad 1 - \rho(\theta^{(j)}, \xi^{(j)}) \end{cases},$$

where

$$\rho(\theta^{(j)}, \xi^{(j)}) = \min\left(\frac{p_n(\xi^{(j)})}{p_n(\theta^{(j)})}, 1\right).$$

Implication of the algorithm is the ergodicity of the chain, that is, the chain satisfies the law of large numbers:

$$\frac{1}{B}\sum_{t=1}^{B} f(\theta^{(t)}) \xrightarrow{p} \int_{\Theta} f(\theta) p_n(\theta) d\theta < \infty.$$

Notes:

1. The parameter σ^2 is regulated such that acceptance rate ρ is about .3-.5. Other parameters can be regulated as well.

2. A good software package have been developed by Charles Geyer. It is available through his page. R also has some new MCMC packages. Of course, it is very easy to code it up, though professional packages offer faster implementations.

3. For more general versions of Metropolis, see references. Also, extensive treatments are available in such references as Casella and Robert's book and Jun Liu's book. Chib's and Geweke's handbook chapters in Handbook of Econometrics are good references.

4. Formal computational complexity for concave L_n (Lovasz and Vempala (2003))

$$O(\dim(\theta)^3),$$

for non-concave L_n (Belloni and Chernozhukov (2006))

$$O(\dim(\theta)^3).$$

The latter holding only in large samples, under the conditions of Bayes CLT.

- Q-Bayes Estimator and Simulated Annealing:

$$\lim_{\lambda \to \infty} \frac{\int_{\Theta} \theta e^{\lambda L_n(\theta)} \pi(\theta)\, d\theta}{\int_{\Theta} e^{\lambda L_n(\theta)} \pi(\theta)\, d\theta} = \operatorname*{argmax}_{\theta \in \Theta} L_n(\theta)$$

The parameter $1/\lambda$ is called temperature.

- The nice part about quasi-Bayesian or Bayesian estimators is that to compute posterior means, no need to send $\lambda \to \infty$.

Part 2. Monte-Carlo Examples

- Simulation Example: Instrumental Quantile Regression

$$Y = D'\beta + u, \quad u = \sigma(D)\epsilon,$$
$$D = \exp N(0, I_3), \quad \epsilon = N(0,1)$$
$$\sigma(D) = (1 + \textstyle\sum_{i=1}^{3} D_{(i)})/5$$

- Instrument moment condition

$$g_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\tau - 1(Y_i \leq \alpha + D'\beta))Z, \text{ where } Z = (1, D).$$

- Weight matrix

$$W = \left[\frac{1}{n}\sum_{i=1}^{n}(\tau(1-\tau))Z_i Z_i'\right]^{-1}.$$

Table 1. Comparison of quasi-bayesian estimators with least absolute deviation estimator (median regression)

Estimator	rmse	mad	mean bias	med. bias	med.ad
n=200					
Q-mean	.0747	.0587	.0174	.0204	.0478
Q-median	.0779	.0608	.0192	.136	.0519
LAD	.0787	.0628	.0067	.0092	0.051
n=800					
Q-mean	.0425	.0323	-.0018	-.0003	0.028
Q-median	.0445	.0339	-.0023	.0001	.0295
LAD	.0498	.0398	.0007	.0025	.0356

Table 2. Comparison of quasi-bayesian inference with standard inference

Inference	**coverage**	**length**
n=200		
Q-equal tail	.943	.377
Q-symmetric(around mean)	.941	.375
QR: HS	.659	.177

Inference	**coverage**	**length**
n=800		
Q-equal tail	.92	.159
Q-symmetric(around mean)	.917	.158
QR: HS	.602	.082

- Simulation Examples: censored regression model

$$Y^* = \beta_0 + X'\beta + u$$
$$X \sim N(0, I_3), \quad u = X_2^2 N(0,1),$$
$$Y = \max(0, Y^*)$$

- Quasi-Bayes estimator to the Powell CQR objective function

$$L_n(\beta) = -\sum_{i=1}^{n} |Y_i - \max\left(0, X_i'\beta\right)|$$

Table 3. Comparison of quasi-bayesian estimators with censored quantile regression estimates obtained using iterated linear programming (100 simulation runs)

Estimator	rmse	mad	mean bias	med. bias
n=400				
Q-posterior-mean	0.473	0.378	0.138	0.134
Q-posterior-median	0.465	0.372	0.131	0.137
Iterated LP(10)	0.518	0.284	0.04	0.016
	3.798	0.827	-0.568	-0.035
n=1600				
Q-posterior-mean	0.155	0.121	-0.018	0.0097
Q-posterior-median	0.155	0.121	-0.02	0.0023
Iterated LP(7)	0.134	0.106	0.04	0.067
	3.547	0.511	0.023	-0.384

Part 3. Empirical Applications

- Dynamic Risk Forecasting, cf. Chernozhukov and Hong (2003),
- Dynamic Games, cf, Ryan
- Complete Information Games, Bajari, Hong, Ryan
- Pricing Kernels, Todorov

Application to Dynamic Risk Forecasting

- Dataset

  Y_t, the one-day returns, the Occidental Petroleum (NYSE:OXY) security,

  X_t, a constant, lagged one-day return of Dow Jones Industrials (DJI), the lagged return on the spot price of oil (NCL, front-month contract on crude oil on NYMEX), and the lagged return Y_{t-1}.

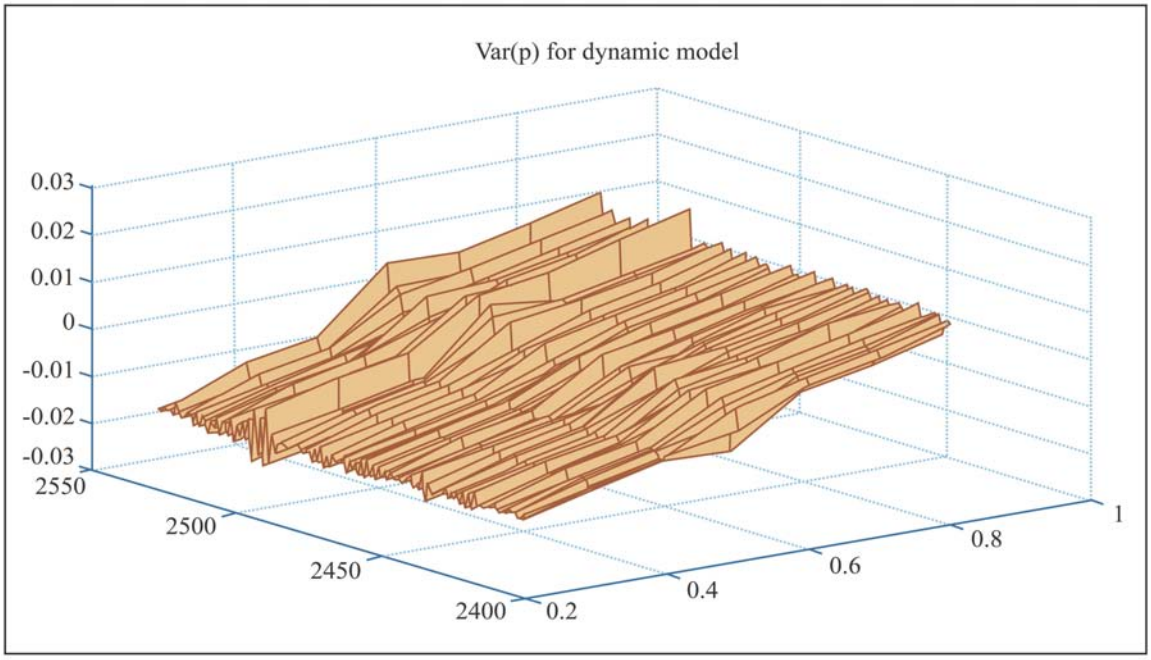
- Conditional Quantile Functions and Estimation

- Linear Model

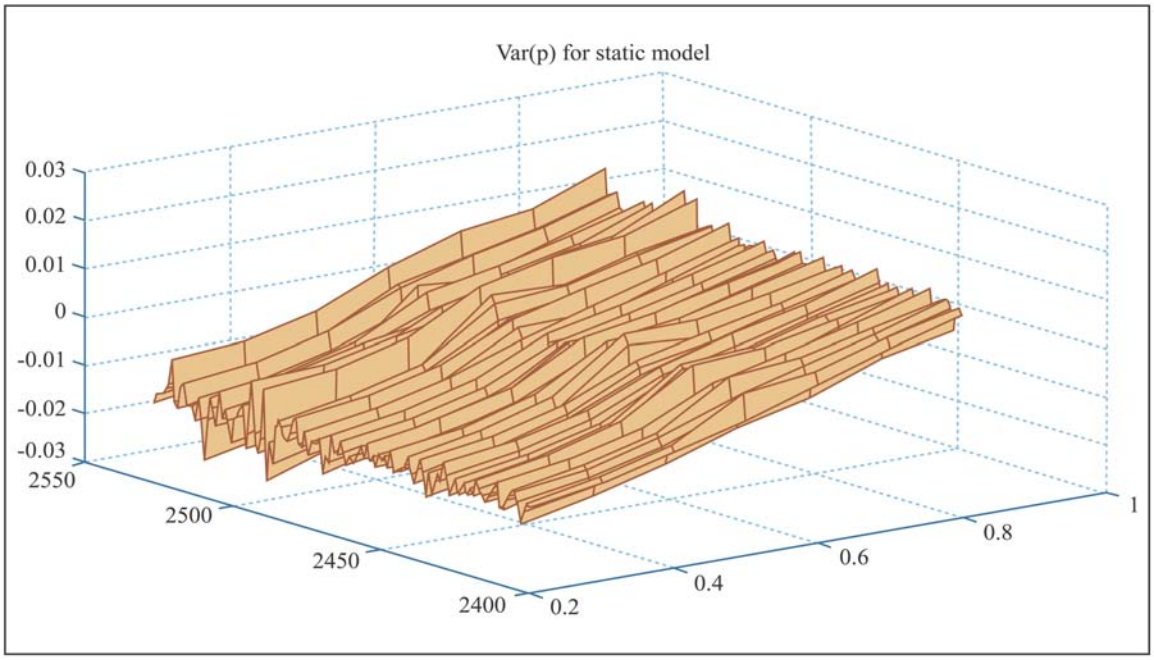
$$q_t(\tau) = X_t' \, \theta(p),$$

- Semi-linear Dynamic Model a-la Engle:

$$q_t(\tau) = X_t'\theta(\tau) + \rho(\tau) q_{t-1}(\tau).$$



Recursive VaR Surface in time-probability space.



Images by MIT OpenCourseWare.

Non-recursive VaR Surface in time-probability space

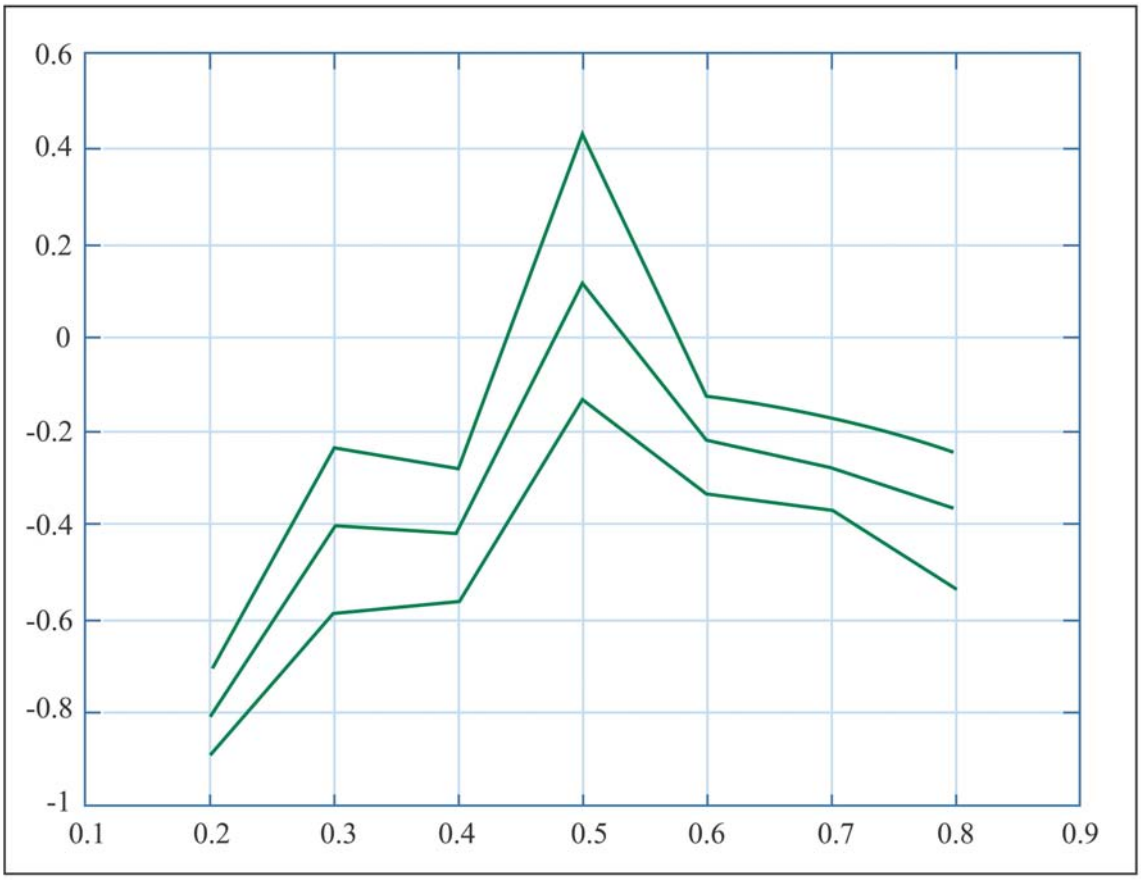
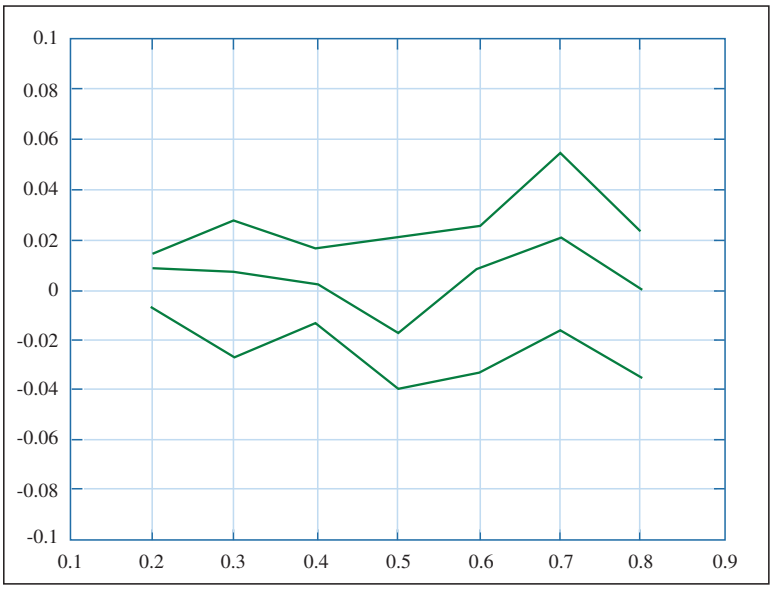
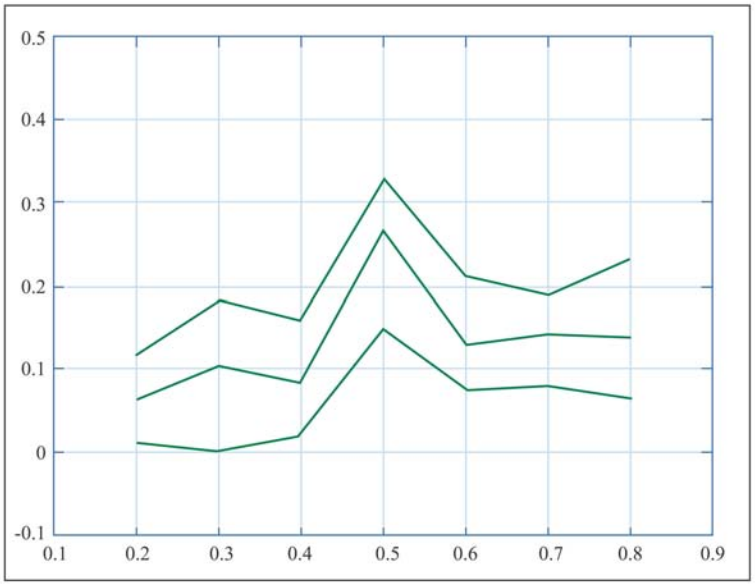


Image by MIT OpenCourseWare.

$\widehat{\varrho}(\tau)$ for $\tau \in [.2, .8]$ and the 90% confidence intervals.



$\widehat{\theta}_2(\tau)$ for $\tau \in [.2, .8]$ and the 90% confidence intervals.



Images by MIT OpenCourseWare.

$\widehat{\theta}_3(\tau)$ for $\tau \in [.2, .8]$ and the 90% confidence intervals.

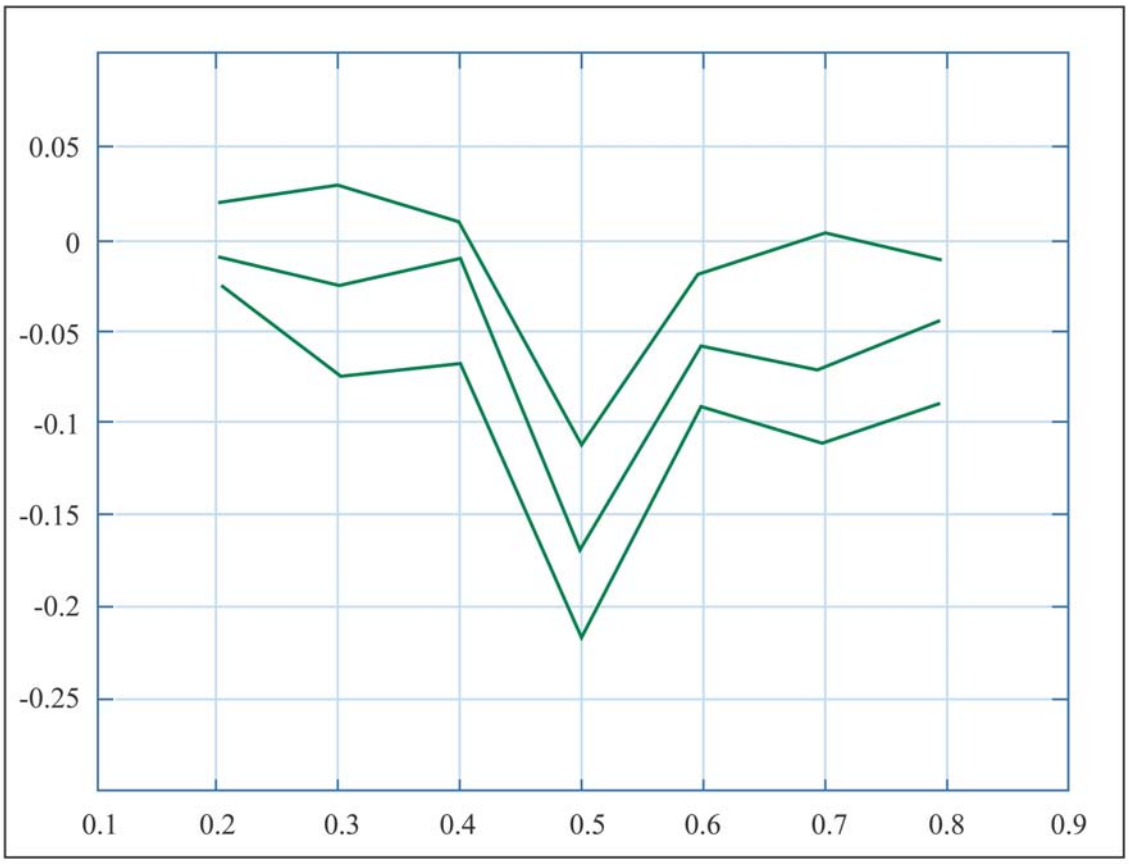


Image by MIT OpenCourseWare.

$\widehat{\theta}_4(\tau)$ for $\tau \in [.2, .8]$ and the 90% confidence intervals.

Part 4. Large Sample Theory of Quasi-Bayesian Estimators – Formal Development

Assumption 1 (Parameter) *θ_0 belongs to the interior of a compact convex subset Θ of $\mathbb{R}^d$.*

Assumption 2 (Identification) *For any $\delta > 0$, there is $\epsilon > 0$:*

$$P\left\{\sup_{|\theta-\theta_0|\geq\delta}\frac{1}{n}\Big(L_n(\theta)-L_n(\theta_0)\Big)\leq-\epsilon\right\}\longrightarrow 1.$$

Assumption 3 (Linear Quadratic Expansion) *For θ in a ball at θ_0,*

i. $L_n(\theta)-L_n(\theta_0)=$

$$(\theta-\theta_0)'\Delta_n(\theta_0)-\tfrac{1}{2}(\theta-\theta_0)'[nJ(\theta_0)](\theta-\theta_0)+R_n(\theta),$$

ii. $\Delta_n(\theta_0)/\sqrt{n}\xrightarrow{d}N(0,\Omega)$,

iii. *Ω and $J(\theta_0)$ are positive-definite, constant matrices,*

iv. *for each $\epsilon>0$ there is sufficiently small $\delta>0$ such that*

$$\limsup_n\ P\left\{\sup_{|\theta-\theta_0|\leq\delta}\frac{|R_n(\theta)|}{1+n|\theta-\theta|^2}>\epsilon\right\}<\epsilon.$$

Comments:

1. Assumptions and proofs are generally patterned but differ from Bickel and Yahav (1969) and Ibragimov and Hasminskii (1981).

2. Differences due to inclusions of non-likelihoods in the framework.

3. Conditions also involve the Huber-style conditions used in exremum analysis.

4. No direct assumptions on sampling mechanisms. Results apply quite generally.

Sensibility of Assumption 4.iv is immediate from usual Cramer-Amemiya restrictions.

Lemma 1 *Assumption 4.iv holds with*

$$\Delta_n(\theta_0) = \nabla_\theta L_n(\theta_0) \text{ and } J(\theta_0) = \nabla_{\theta\theta'} M(\theta_0),$$

if for $\delta > 0$, $L_n(\theta)$ *is twice differentiable in* θ *when* $|\theta - \theta_0| < \delta$

$$\nabla_\theta L_n(\theta_0)/\sqrt{n} \xrightarrow{d} N(0, \Omega)$$

and for each $\epsilon > 0$,

$$P\Big(\sup_{|\theta-\theta_0|<\delta} \Big| \nabla_{\theta\theta'} L_n(\theta)/n - \nabla_{\theta\theta'} M(\theta) \Big| > \epsilon \Big) \longrightarrow 0$$

where $M(\theta)$ *is twice continuously differentiable at* θ_0.

Asymptotic Results

Using the obtained earlier beliefs

$$p_n(\theta) = \frac{e^{L_n(\theta)}\pi(\theta)}{\int_\Theta e^{L_n(\theta)}\pi(\theta)d\theta},$$

for large n, the belief $p_n(\theta)$ is approximately a **random normal density** with

$$\text{random mean} = \theta^* \approx \theta_0 - J(\theta_0)^{-1}\Delta_n(\theta_0)/\sqrt{n}$$

and constant *variance* parameter

$$\text{variance} = \ J(\theta_0)^{-1}/n.$$

Intuition for this result is simple: Define the local parameter

$$u = \sqrt{n}(\theta - \theta_0)$$

and also the local parameter relative to the (first order approximation to) extremum estimator

$$h = \sqrt{n}(\theta - \theta^*) = u - J(\theta_0)^{-1}\Delta_n(\theta_0).$$

The quasi-posterior belief about u is

$$\bar{p}_n(u) = \frac{1}{\sqrt{n^d}}p_n(\theta_0 + u/\sqrt{n})$$

and about h

$$p_n^*(h) = \bar{p}_n(h + J(\theta_0)^{-1}\Delta_n(\theta_0)/\sqrt{n}).$$

Then,

$$\begin{aligned}
\bar{p}_n(u) &\propto e^{\ L_n(\theta_0+u/\sqrt{n})-L_n(\theta_0)} \\
&\propto e^{u'\Delta_n(\theta_0)/\sqrt{n}-\frac{1}{2}u'J(\theta_0)u} \cdot (1+o_p(1)) \\
&\propto e^{-\frac{1}{2}(u-J(\theta_0)^{-1}\Delta_n(\theta_0)/\sqrt{n})'J(\theta_0)(u-J(\theta_0)^{-1}\Delta_n(\theta_0))}(1+o_p(1))
\end{aligned}$$

Hence

$$p_n(h) \propto e^{-\frac{1}{2}\ h'J(\theta_0)h} \cdot (1+o_p(1)).$$

Theorem 1 (Beliefs in Large Sample) *In large samples, under Assumptions 1-3 + $\pi(\theta)$ continuous and positive on Θ,*

$$p_n^*(h) \approx p_\infty^*(h) = \frac{\sqrt{\det J(\theta_0)}}{\sqrt{(2\pi)^d}} \cdot e^{-\frac{1}{2}h'J(\theta_0)h}.$$

where $\approx$ means that, for any $\alpha \geq 0$,

$$TVM = \int_{H_n} \left(1+|h|^\alpha\right)\left|p_n^*(h) - p_\infty^*(h)\right|dh \xrightarrow{p} 0.$$

Theorem 2 (QBE in Large Sample) *Under assumption 1-3, for symmetric convex penalty functions ℓ and conditions on the prior as in the previous theorem*

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \sqrt{n}(\theta^* - \theta_0) + o_p(1) = Z_n + o_p(1)$$

where

$$Z_n = J(\theta_0)^{-1}\Delta_n/\sqrt{n},$$

and

$$Z_n \xrightarrow{d} \mathcal{N}(0, J(\theta_0)^{-1}\Omega(\theta_0)J(\theta_0)^{-1}).$$

If

$$\Omega(\theta_0) = J(\theta_0) \qquad (*)$$

then quasi-posterior quantiles are valid for classical inference

(*) is a **generalized information equality**

(*) holds for GMM with optimal weight matrix, minimum distance, empirical likelihood, and properly weighted regression objective functions.

(*) holds when $L_n(\theta)$ is the log-likelihood function that satisfied information equality

Suppose we want to do inference about a real function of the parameter

$$g(\theta_0),$$

and g is continuously differentiable at θ_0. For example, $g(\theta_0)$ can be the j-th component of θ_0.

Define

$$F_{g,n}(x) = \int_{\Theta} 1\{g(\theta) \leq x\} p_n(\theta) d\theta.$$

and

$$c_{g,n}(\alpha) = \inf\{x : F_{g,n}(x) \geq \alpha\}.$$

Here $c_{g,n}(\alpha)$ is our posterior α-quantile, and $F_{g,n}(x)$ is the posterior cumulative distribution function.

Then a posterior CI is given by

$$[c_{g,n}(\alpha/2), c_{g,n}(1-\alpha/2)].$$

These CI's can be computed by using the $\alpha/2$ and $1-\alpha/2$ quantiles of the MCMC sequence

$$(g(\theta^{(1)}), ..., g(\theta^{(B)}))$$

and thus are quite simple in practice.

The usual Δ- method intervals are of the form

$$[g(\widehat{\theta}_e) + q_{\alpha/2}\frac{\sqrt{\widehat{\nabla_\theta g(\theta)}' J_n(\theta_0)^{-1} \widehat{\nabla_\theta g(\theta)}}}{\sqrt{n}}, \quad g(\widehat{\theta}_e) + q_{1-\alpha/2}\frac{\sqrt{\widehat{\nabla_\theta g(\theta)}' \widehat{J_n(\theta_0)^{-1}} \widehat{\nabla_\theta g(\theta)}}}{\sqrt{n}}],$$

where q_α is the α-quantile of the standard normal distribution.

Theorem 3 (Large Sample Inference I) *Suppose Assumptions 1-4 and (*) hold. Then for any* $\alpha \in (0,1)$

$$c_{g,n}(\alpha) - g(\widehat{\theta}) - q_\alpha \frac{\sqrt{\nabla_\theta g(\theta_0)' J_n(\theta_0)^{-1} \nabla_\theta g(\theta_0)}}{\sqrt{n}} = o_p(\frac{1}{\sqrt{n}}),$$

and

$$\lim_{n\to\infty} P^*\Big\{c_{g,n}(\alpha/2) \le g(\theta_0) \le c_{g,n}(1-\alpha/2)\Big\} = 1-\alpha.$$

Can also use the Quasi-posterior variance as an estimate of the inverse of the population Hessian matrix $J_n^{-1}(\theta_0)$, and combine it with any available estimate of $\Omega_n(\theta_0)$ (which typically is easier to obtain) in order to obtain the Δ-method style intervals.

Theorem 4 (Large Sample Inference II) *Suppose Assumptions 1-4 hold. Define for* $\widehat{\theta} = \int_\Theta \theta p_n(\theta) d\theta$,

$$\widehat{J}_n^{-1}(\theta_0) \equiv \int_\Theta n(\theta - \widehat{\theta})(\theta - \widehat{\theta})' p_n(\theta) d\theta,$$

and

$$c_{g,n}(\alpha) \equiv g(\widehat{\theta}) + q_\alpha \cdot \frac{\sqrt{\widehat{\nabla_\theta g(\theta)}' \widehat{J_n(\theta_0)^{-1}} \widehat{\Omega_n(\theta_0)} \widehat{J_n(\theta_0)^{-1}} \widehat{\nabla_\theta g(\theta)}}}{\sqrt{n}},$$

where $\widehat{\Omega}_n(\theta_0) \xrightarrow{p} \Omega(\theta_0)$. *Then* $\widehat{J}_n(\theta_0)^{-1} \xrightarrow{p} J_n(\theta_0)^{-1}$, *and*

$$\lim_{n\to\infty} P^*\Big\{ c_{g,n}(\alpha/2) \leq g(\theta_0) \leq c_{g,n}(1-\alpha/2) \Big\} = 1 - \alpha.$$

• In practice $\widehat{J}_n(\theta_0)^{-1}$ is computed by multiplying by n the variance-covariance matrix of the MCMC sequence $S = \left(\theta^{(1)}, \theta^{(2)}, ..., \theta^{(B)}\right)$.

Conclusions:

- Generic Computability using Markov Chain Monte-Carlo. Quasi-Bayesian estimators are relatively easy to compute by drawing a sample whose marginal distribution is p_n.

  " Replace optimization with integration and integration is cheap and numerically stable while optimization is neither" (Heckman)

- Theoretical framework unifies both Bayesian and Non-bayesian – but similarly defined – estimators.

- Quasi-Bayesian Estimator have good formal properties; they are as good as extremum estimators.