

Entropy-based econometric techniques:
A Review of Recent Research

Yuichi Kitamura
Department of Economics
Yale University

This talk:

- Frequentist Approach

 - Statistical Efficiency

 - Robustness

 - Computation

- Bayesian Approach

 - Bayesian semiparametrics

- Computational Issues

Common theme:

Use of relative entropy $D(P|Q) = \int \log(dP/dQ)dP,$

P, Q: probability measures

(or its generalization: $\int \phi(dP/dQ)dQ,$ where ϕ is convex)

Parametric modeling and relative entropy

- Close connection between likelihood and relative entropy

White (1982)

Vuong (1989)

— ML as minimum relative entropy procedure

- Information theoretic results on optimality of ML

Hoeffding (1963)

Bahadur (1960, 1961)

Recent literature

- Relative entropy provides a practical and useful algorithm when model is not completely specified (*semiparametric models*)
- Optimality — in terms of *efficiency* or *robustness*
- See the Golan-Judge-Miller monograph for comprehensive overview
- A leading example: moment restriction model

Suppose

$$E[g(\mathbf{x}, \theta)] = \int g(\mathbf{x}, \theta) dP_0 = 0, \theta \in \Theta$$

identifies θ_0 uniquely.

Relative entropy in moment restriction model

- Define $v(\theta) := \min_{\mathbf{P}} D(\mathbf{P}_0 | \mathbf{P})$ s.t. $\int \mathbf{g}(\mathbf{x}, \theta) d\mathbf{P} = 0$.
- Then $\theta_0 = \arg \min_{\theta \in \Theta} v(\theta)$.
- But by Fenchel duality,

$$v(\theta) = \max_{\boldsymbol{\gamma} \in \mathbf{R}^q} - \int \log(\mathbf{1} + \boldsymbol{\gamma}' \mathbf{g}(\mathbf{x}, \theta)) d\mathbf{P}_0$$

- This suggests:

$$\hat{\theta} = \arg \min_{\theta} \max_{\boldsymbol{\gamma}} - \frac{1}{n} \sum_{i=1}^n \log(\mathbf{1} + \boldsymbol{\gamma}' \mathbf{g}(\mathbf{x}_i, \theta))$$

= EL (Empirical Likelihood) estimator

\Rightarrow **EL estimator as minimum entropy estimator.**

Exponential Tilting

If, instead, we solve $\min_{\mathbf{P}} D(\mathbf{P}|\mathbf{P}_0)$ s.t. $\int g(\mathbf{x}, \theta) d\mathbf{P} = 0$, we get

$$\hat{\theta} = \arg \min_{\theta} \max_{\gamma} -\frac{1}{n} \sum_{i=1}^n \exp(\gamma' g(\mathbf{x}_i, \theta))$$

= Exponential tilting estimator (Kitamura and Stutzer 1997)

Optimality of EL in terms of “Large Deviations”

Kitamura (2001), Kitamura, Santos and Sheikh (2008) [Overidentification Restrictions Test]

Kitamura and Otsu (2006) [Parameter Estimation, Parametric Test]

Caney (2008) [Inference with Moment Inequality]

- EL yields “efficient” procedures in terms of large deviations

Motivating Relative entropy: Sanov's Theorem

- Large Deviation Principle (LDP) for Empirical Distribution

Setup

$$X_i \sim_{\text{iid}} P_0$$

P_n : Empirical distribution

\mathbf{M} : the space of all probability distribution functions equipped with the weak topology

$$\mathcal{P} \subset \mathbf{M}, \quad \mathcal{P}^\circ: \text{interior of } \mathcal{P}, \quad \bar{\mathcal{P}}: \text{closure of } \mathcal{P},$$

Sanov's Theorem

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \Pr(\mathbf{P}_n \in \mathcal{P}^\circ) \geq - \inf_{P \in \mathcal{P}^\circ} D(P, P_0)$$
$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \Pr(\mathbf{P}_n \in \bar{\mathcal{P}}) \leq - \inf_{P \in \bar{\mathcal{P}}} D(P, P_0)$$

- The factor on the RHS is generally called the rate function.
- The rate function for the LDP for empirical distribution is given by the relative entropy.

Lemma (Varadhan)

- $f(\cdot)$: Lower semicontinuous and non-negative functional of $P \in \mathbf{M}$

Then

$$\liminf_{n \rightarrow \infty} \left(\mathbb{E}_{P_0} [f(P_n)] \right)^{1/n} \geq \sup_P f(P) e^{-D(P|P_0)}.$$

Optimal Robustness

Define an alternative “entropy”

$$H(P|Q) = \left\{ \int (dP^{1/2} - dQ^{1/2})^2 \right\}^{1/2} \text{ (i.e. Hellinger distance)}$$

$$v_H(\theta) := \min_P D(P|Q) \quad \text{s.t.} \quad \int g(x, \theta) dP = 0$$

Then again by Fenchel duality,

$$\theta_0 = \arg \min_{\theta} \max_{\gamma} - \int \frac{1}{1 + \gamma' g(z, \theta)} dP_0$$

So define:

$$\hat{\theta}_H = \arg \min_{\theta} \max_{\gamma} - \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + \gamma' g(z_i, \theta)}.$$

Result (Kitamura, Otsu and Evdokimov, 2009)

$\hat{\theta}_H$ is optimally robust (robustness against small perturbation due to data contamination): it minimizes

$$\text{AMSE} = \lim_{M \rightarrow \infty} \liminf_{n \rightarrow \infty} \sup_{P \in B_h(P_0, \delta/\sqrt{n})} \int (M \wedge (\hat{\theta} - \theta_0)^2) dP_0^{\otimes n}.$$

for arbitrary $\delta > 0$ with $B_h(P, \delta) = \{Q : H(Q|P) \leq \delta\}$

Computation and Implementation

Take exponential tilting with moment function $g(x, \theta) \in \mathbf{R}^q$

(P) “Primal problem”

$$\min_{p_i, i=1, \dots, n, \theta \in \Theta} D(P_n | \mu_n) \text{ s.t. } \int g(x, \theta) dP_n = 0$$

$$\mu_n := \text{empirical measure} \quad P_n := \sum_{i=1}^n p_i \delta_{x_i}$$

(D) “Dual problem”

$$\min_{\theta \in \Theta} \max_{\gamma \in \mathbf{R}^q} -\frac{1}{n} \sum_{i=1}^n \exp(\gamma' g(x_i, \theta))$$

(**P**) might seem infeasible, but recent advances numerical methods with sparse matrices can work

cf. Conlon (2008), Zedlewski (2008)

- Good news:

Converges very fast (when it does)

A “canned” program works (when it does)

- Bad news:

Some report problems when g is highly nonlinear

(**D**) has been extensively studied.

- Good news:

Seems to be more stable when g is nonlinear

“Inner loop” is trivial to solve

- Bad news:

Nested algorithm

Careful programming is crucially important

Bayesian Analysis without fully specified likelihood function

- Bayesian Method of Moments (BMOM) (Zellner 1996, Zellner and Tobias 2001)
- Bayesian “exponential tilted empirical likelihood” (Schenach, 2005)
- Bayesian limited information inference (Kwan 1998, Kim 2002)

Bayesian Nonparametrics/Semiparametrics

- Rapidly growing literature (cf. Ghosh and Ramamoorthi (2003))
- Prior on infinite dimensional spaces
- Some cautionary tales

Diaconis and Freedman, Cox

- Positive results

Barron, Schervish and Wasserman (1992)

Shen (2002)

Many recent results by Ghosal, van der Vaart

Fast algorithm by Escobar and West

Kitamura and Otsu (in progress)

- Bayesian semiparametrics in moment condition model

$$E[g(x, \theta)] = \int g dP = 0$$

- Want to incorporate

Prior information on smoothness of P (nonparametric)

Prior information on θ (parametric)

— in possibly overidentified models

Issues

- (1) Incorporating smoothness
- (2) Naive use of Dirichlet Process prior (or other standard non-parametric) prior fails to satisfy the model restriction

Our solution

- (1) Dirichlet Process Mixture (DPM)
- (2) Use exponential tilting — relative entropy minimization in the space of priors.

Kitamura and Otsu (2009)

Consistency, Rate of convergence

Asymptotic normality of posterior

(Bernstein von Mises Theorem)

- This framework should extend to moment inequality models.

Topics for further investigation

Further results on

- (1) Robustness, including testing
- (2) Treatment of moment inequality models
- (3) Use of Bayesian nonparametric/semiparametric techniques.
- (4) Computational algorithm

would be useful.