Applications of Asymptotic Statistical Decision Theory in Econometrics

Keisuke Hirano
University of Arizona

Jack Porter
University of Wisconsin

April 2011

Asymptotic Statistical Decision Theory

A perspective, and also a toolkit.

Perspective:

- Specify possible probability distributions, possible actions, and consequences of actions
- Evaluate procedures by their performance under "all possible" probability distributions.

Toolkit:

- Simplification of statistical *problems*
- Use results from classic statistical decision theory in new contexts.

Outline

- Wald Framework
- Le Cam's Approach and the Asymptotic Representation Theorem
- Nondifferentiable Functionals
- Treatment Rules

Wald Framework

- Parameter space: Θ
- Observe a random variable Z whose distribution is determined by $\theta \in \Theta$.
- Statistical Experiment: $\{P_\theta : \theta \in \Theta\}$.
- Decision Problem characterized by an action space $\mathcal{A}$ and loss function $L(\theta, a) : \Theta \times \mathcal{A} \to \mathbb{R}$.

- A decision procedure $\delta(z)$ maps realizations of Z into actions.
- Risk of procedure δ:

$$R(\theta, \delta) = E_\theta \left[L(\theta, \delta(Z)) \right] = \int L(\theta, \delta(z)) dP_\theta(z).$$

- We usually extend these definitions to allow for auxiliary randomization (e.g. $\delta(z, u)$ where u is independent random variable).

In general, risk functions for different procedures cross. Can rank rules by:

- Bayes risk: let Π be a prob. measure on θ, and define
$$r(\Pi, \delta) = \int R(\theta, \delta) d\Pi(\theta).$$
- Minmax risk: evaluate δ by
$$\sup_{\theta \in \Theta} R(\theta, \delta).$$
- Minmax regret, Gamma-minmax, etc.

Example: Normal Shift

Let $\Theta = \mathbb{R}$, and let P_θ be the $N(\theta, 1)$ distribution, i.e. the model is

$$Z \sim N(\theta, 1).$$

Point estimation: $\mathcal{A} = \Theta = \mathbb{R}$, and we interpret $d(z)$ as a point estimate of θ.

Squared Error Loss: $L(\theta, a) = (\theta - a)^2$.

What is the "best" point estimator of θ?

- Notice that $\{P_\theta : \theta \in \Theta\}$ has an additive shift (translation) structure.
- Consider equivariant estimators:

$$\delta(z + c) = \delta(z) + c.$$

$$\Rightarrow \delta(z) = \delta(0) + z.$$

- Hence every equivariant estimator must have the form $\delta(z) = z + b$ for some constant b.

- By the Hunt-Stein theorem, we can restrict attention to equivariant estimators.
- Moreover, equivariant estimators have constant risk over θ.
- So it is easy to solve for "best" estimator.
- In this example, $\delta(z) = z$ is minmax, MLE, and flat-prior Bayes.

Similar arguments can be used for:

- Other translation-equivariant models, e.g. multivariate normal with known variance matrix, exponential shift.
- Other loss functions that have a translation form: $L(\theta, a) = f(\theta - a)$.
- Randomized estimators.

On the other hand, most econometric models are not translation-equivariant and are too complex to derive exact optimality results.

Le Cam Framework

Consider a finite-dimensional parametric model:

$$Z_1, Z_2, \ldots \overset{\text{iid}}{\sim} P_\theta,$$

where

$$\theta \in \Theta \subset \mathbb{R}^k.$$

For given $n \geq 1$, can view data as $Z^{(n)} = (Z_1, \ldots, Z_n)$ and statistical model as

$$\mathcal{E}_n = \{P_\theta^n : \theta \in \Theta\}.$$

This defines a sequence of statistical experiments.

Suppose $\{\mathcal{E}_n\}$ satisfies the usual (DQM) smoothness condition.

As $n \to \infty$, the parameter θ can be "learned" perfectly.

Local reparametrization: fix θ_0, consider sequences of parameters $\theta_0 + h/\sqrt{n}$ for $h \in \mathbb{R}^k$.

The $1/\sqrt{n}$ rate has the property that the likelihood ratio

$$\frac{dP_{\theta_0 + h/\sqrt{n}}}{dP_{\theta_0}}(Z^{(n)})$$

has nondegenerate limit.

Let J_0 denote Fisher information for the parametric model at θ_0. (assume nonsingular)

Let δ_n be a (normalized) sequence of statistics/decision rules satisfying:

$$\delta_n \overset{h}{\rightsquigarrow} L_h, \tag{1}$$

where $\overset{h}{\rightsquigarrow}$ indicates weak convergence under $\theta_0 + h/\sqrt{n}$ to laws L_h.

Example:

$$\delta_n = \sqrt{n}(\hat{\theta}_{ML} - \theta_0).$$

Asymptotic Representation Theorem

Van der Vaart (1991): for any rule δ_n satisfying (1), there exists a randomized rule $\delta^*(Z, U)$, where $Z \sim N\left(h, J_0^{-1}\right)$ such that

$$L_h = \mathcal{L}_h\left[\delta^*(Z, U)\right],$$

Limit distributions of δ_n are matched by the exact distributions of some rule (δ^*) in the $N(h, J_0^{-1})$ model.

$N(h, J_0^{-1})$ is the **limit experiment**. (Le Cam)

This characterizes limits of every possible decision rule, and suggests to “solve” the problem in the normal case.

Typically, it is possible to construct matching sequences of estimators in the original experiment:

- MLE in original experiment will be matched by MLE in limit experiment.
- Bayes (with smooth prior) will be matched by flat-prior Bayes in the limit experiment.
- For point estimation, this gives optimality of MLE and Bayes.

- The ART can be extended to semiparametric problems where Θ is infinite-dimensional, and to some nonregular models.
- There is a different, nonlocal approximation theory for nonparametric regression and density estimation (e.g. Nussbaum, Brown and Low).

Application: Nondifferentiable Functionals

Consider

$$Y_i \overset{\text{iid}}{\sim} G_\theta, \quad \theta \in \Theta \subset \mathbb{R}^k.$$

Object of interest:

$$\kappa(\theta) = \min\{\theta_1, \ldots, \theta_k\}.$$

This type of problem arises in bounds/moment inequality models (quantity of interest is bounded above by two features of the data).

For example:

- Manski-Pepper (2000): bounds on treatment effects defined by maxima and minima of conditional expectations.
- Haile-Tamer(2003): $G_m(b)$ is CDF for bids in auctions with m bidders. If bidders bid no more than their valuation, valuation CDF satisfies

$$F(v) \leq \min_m G_m(v).$$

- Andrew-Soares (2007): general inference procedures for moment inequality models.

Typically there exist estimators $\hat{\theta}_n$ satisfying

$$\sqrt{n}(\hat{\theta}_n - \theta_0 - h/\sqrt{n}) \overset{h}{\rightsquigarrow} N(0, J_0^{-1})$$

(asymptotically unbiased, local asymptotic minmax).

But

$$\hat{\kappa}_n := \kappa(\hat{\theta}_n)$$

will be biased (downwards) due to min operator.

Can bias be removed? What is optimal estimator?

Le Cam framework suggests to consider the limit experiment:

$$Z \sim N(h, J_0^{-1}), \quad \text{where } J_0 \text{ is known.}$$

Let $\delta(Z)$ be an estimator of $\kappa(h)$.

Argument adapted from Blumenthal and Cohen (1968).

Suppose $\delta(Z)$ is unbiased:

$$E_h[\delta(Z)] = \kappa(h) \quad \forall h \in \mathbb{R}^k.$$

Write

$$E_h[\delta(Z)] = \int \delta(z) dN(z|h, J_0^{-1}) dz.$$

Unbiasedness condition implies validity of differentiating under the integral sign, so:

$$\frac{\partial}{\partial h_1} E_h[\delta(Z)] = \int -(z_1 - h_1)\delta(z) dN(z|h, J_0^{-1}) dz,$$

and LHS is well defined.

Since $E_h[\delta(Z)] = \kappa(h)$, $\kappa(h)$ must be differentiable in h_1.

Unbiasedness condition implies validity of differentiating under the integral sign, so:

$$\frac{\partial}{\partial h_1} E_h[\delta(Z)] = \int -(z_1 - h_1)\delta(z) dN(z|h, J_0^{-1}) dz,$$

and LHS is well defined.

Since $E_h[\delta(Z)] = \kappa(h)$, $\kappa(h)$ must be differentiable in h_1.

Contradiction: $\kappa(h)$ is not differentiable at $h_1 = h_2$. $\square$

Hence in the limit experiment there exists no unbiased estimator for $\kappa(h)$.

Can also show:

- Reducing bias will eventually cause variance to diverge.
- There exists no estimator satisfying the (general) equivariance condition:

$$\delta(z) - \kappa(h) \overset{h}{\sim} F, \quad \forall h,$$

where F does not depend on h.

Another classic result (Fraser, 1952): for $J_0 = I_k$, under mild conditions on δ, the following cannot hold:

$$P_h \left[\delta(Z) \geq \kappa(h)\right] = \beta \quad \forall h.$$

So no median unbiased estimator ($\beta = .5$).

Back to the original problem:

We are doing asymptotics under local sequences of parameters $\theta_0 + h/\sqrt{n}$.

Value of θ_0 will matter.

If $\theta_0 = (\gamma, \gamma, \ldots, \gamma)$, then

$$\kappa(\theta_0 + h/\sqrt{n}) = \theta_0 + \frac{1}{\sqrt{n}}\kappa(h_1, \ldots, h_k).$$

Consider estimator $\hat{\kappa}_n$ with

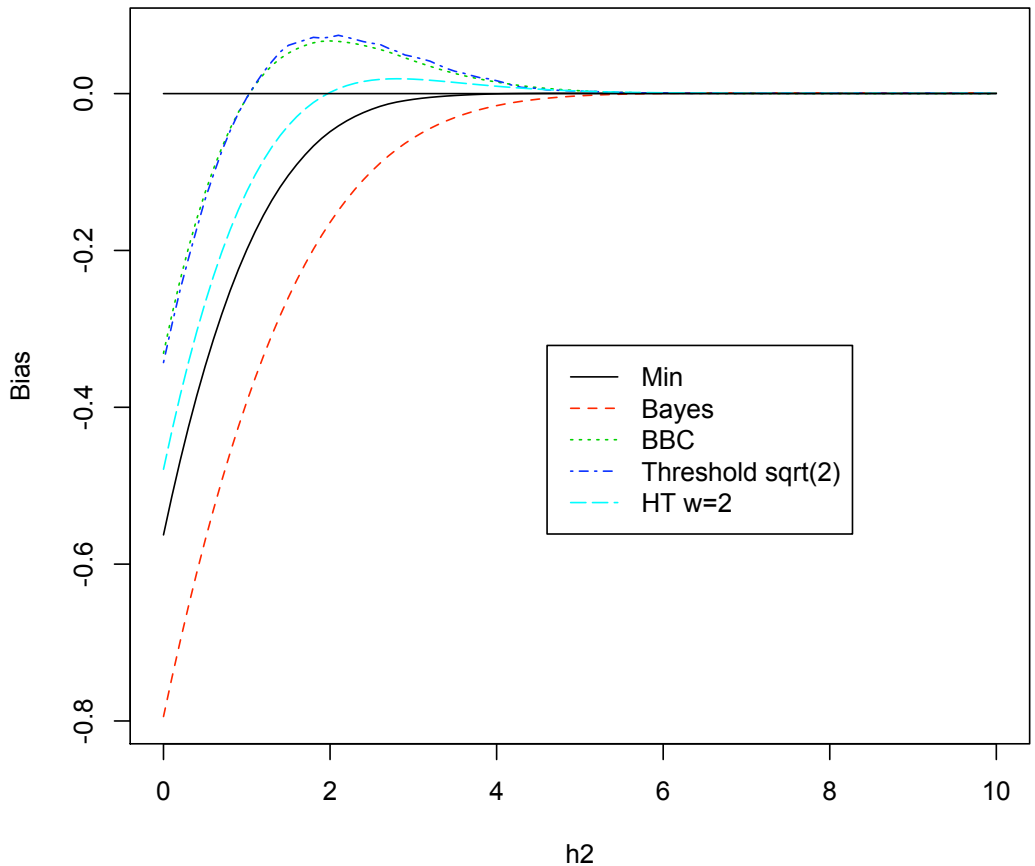
$$\sqrt{n}(\hat{\kappa}_n - \kappa(\theta_0 + h/\sqrt{n})) \overset{h}{\rightsquigarrow} L_h,$$

- Locally unbiased estimator: mean of L_h equals 0 for all h.
- Regular estimator: L_h does not depend on h. (e.g. $N(0, \Sigma)$).
- Locally median unbiased: L_h has median 0.

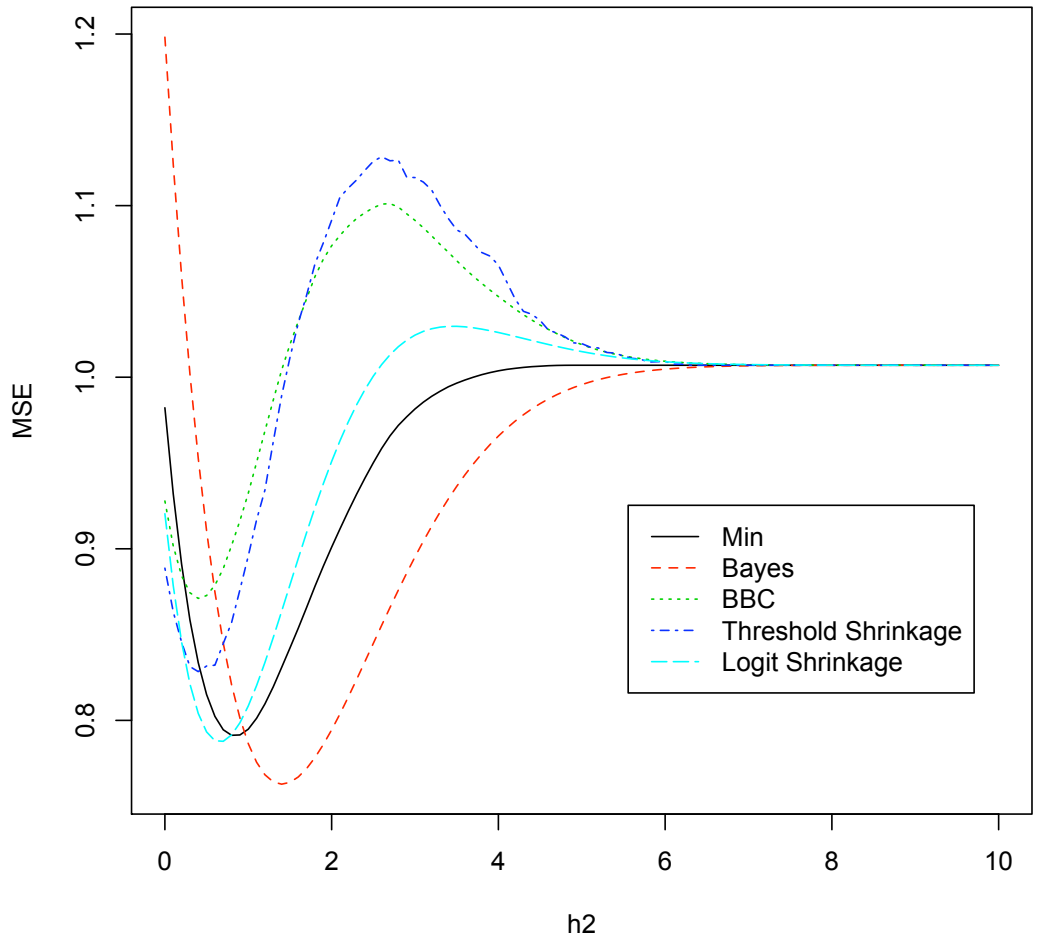
- By ART, limit distributions L_h are matched by distributions of some feasible estimator in the $N(h, J_0^{-1})$ experiment.
- But we know there exist no exactly unbiased or equivariant estimators in the multivariate normal model, and no median unbiased estimators under some restrictions.
- Hence, no asymptotically unbiased, regular, or median unbiased estimators.

- On the other hand, can study the limiting version of the problem, and compare procedures.
- Standard optimality results do not hold, but some minmax results available (Blumenthal and Cohen, Song).
- Can also numerically study the limit experiment:

Bias for $h_1 = 0$, $h_2 \geq 0$



MSE


MSE
1.2
1.1
1.0
0.9
0.8
0
2
4
6
8
10
h2
Min
Bayes
BBC
Threshold Shrinkage
Logit Shrinkage

Application: Treatment Assignment Rules

- Treatment: $T = 0, 1$ (e.g. job training vs. wage subsidy).
- Potential Outcomes: Y_0, Y_1, with distributions $F_0(\cdot|\theta)$, $F_1(\cdot|\theta)$.
- Social Planner chooses a treatment rule

$$\delta = Pr(T = 1).$$

(Could make this conditional on observed individual's characteristics X.)

- Social welfare under treatments 0,1:

$$w_0(\theta), w_1(\theta).$$

- Welfare contrast: $g(\theta) = w_1(\theta) - w_0(\theta)$. Assume continuously differentiable.

- Manski, Dehejia: view this as a statistical problem.
- There is data $Z^{(n)}$ that is informative about θ:

$$Z^{(n)} \sim P_\theta^n.$$

- Example: $Z^{(n)}$ obtained from a randomized evaluation.
- *Statistical* treatment rule:

$$\delta_n(z^{(n)}) = Pr(T = 1 | Z^{(n)} = z^{(n)}).$$

- There is some finite-sample theory (Stoye) but difficult in many settings.
- We consider asymptotic approximations.
- As before, consider sequences

$$\theta_0 + h/\sqrt{n}.$$

- Under regularity conditions, will exist consistent, asymptotically normal estimators for θ.

- If, say, $g(\theta_0) > 0$, then for all h, treatment 1 is (eventually) better and we can learn this.
- But if $g(\theta_0) = 0$, then

$$\sqrt{n} g(\theta_0 + h/\sqrt{n}) \to \dot{g}' h,$$

  and in the limit, we cannot learn perfectly which treatment is better.
- So localizing at θ_0 s.t. $g(\theta_0) = 0$ gives a nontrivial limiting decision problem.

Consider a sequence of treatment rules $\{\delta_n\}$.

Let

$$\beta_n(h) = E_{\theta_0 + h/\sqrt{n}}\left[\delta_n(Z^{(n)})\right].$$

If $\exists \beta(\cdot)$ s.t.

$$\beta_n(h) \to \beta(h) \quad \forall h,$$

Then by ART, exists a rule δ^* in the simple experiment $Z \sim N(h, J_0^{-1})$ s.t.

$$\beta(h) = E_h\left[\delta^*(Z)\right] = \int \delta^*(z) dN(z|h, J_0^{-1}).$$

The limiting version of the problem is one where we observe

$$Z \sim N(h, J_0^{-1}),$$

and have a binary choice where payoffs are based on $\dot{g}'h$.

It can be shown that for many loss functions of interest, can restrict attention to rules of the form

$$\delta(z) = 1(\dot{g}'Z > c)$$

Argument works along one-dimensional subspaces for h and uses Neyman-Pearson Lemma.

Easy to solve for minmax rule: optimization over a scalar c.

Also easy to solve for flat-prior Bayes.

Hirano and Porter (2009): local asymptotic minmax and Bayes-optimal rules, also for semiparametric versions of the problem, generally of the form

$$1(g(\hat{\theta}_n) > c_n),$$

where $\hat{\theta}$ is an efficient estimator.

Extension: Dynamic Treatment Rules (incomplete)

- Murphy (2003), Robins (2004): how to construct rules for assigning treatments over time.
- Example: modify drug dosages based on intermediate outcome measures.
- Use data (from other subjects) to construct dynamic rules.
- But focus so far is on estimating (infeasible) optimal rules, not viewing as a statistical decision problem.

A simple two-period version:

$$T_1 \to X \to T_2 \to Y.$$

- T_1, T_2: binary treatments.
- X: intermediate outcome (assume discrete).
- Y: final outcome.

Suppose

$$X|T_1 = t_1 \sim F_X(\cdot|\theta, t_1),$$

$$Y|T_1 = t_1, X = x, T_2 = t_2 \sim F_Y(\cdot|\theta, t_1, x, t_2),$$

and there is data

$$Z^{(n)} \sim P_\theta^n.$$

Dynamic statistical treatment rule: a pair $\delta_n = (\delta_{n1}, \delta_{n2})$, where

$$\delta_{n1}(z^n) = Pr(T_1 = 1|Z^n = z^n);$$

$$\delta_{n2}(z^n, t_1, x) = Pr(T_2 = 1|Z^n = z^n, t_1, x).$$

Suppose that for every h

$$\beta_{n1}(h) = E_{\theta_0 + h/\sqrt{n}}[\delta_{n1}(Z^n)] \to \beta_1(h)$$
$$\beta_{n2}(h; t_1, x) = E_{\theta_0 + h/\sqrt{n}}[\delta_{n2}(Z^n, t_1, x)] \to \beta_2(h; t_1, x)$$

By ART, there exists a joint rule

$$\delta^*(Z) = (\delta_1^*(Z), \delta_2^*(Z; t_1, x))$$

such that when $Z \sim N(h, J_0^{-1})$, the distribution of $\delta^*(Z)$ is the same as the limiting distribution of δ_n.

In period 2, for a given t_1, x, treatment decision is same as previous single-period problem.

Can use previous results to get optimal δ_2^* and associated risk.

It remains to work out δ_1^*.

Conclusion

- Asymptotic statistical decision theory: an approximation technique for *models* and *statistical decision problems*.
- Still need to solve decision problem in the limit experiment; sometimes much easier.
- Can also be used for nonregular experiments: limit experiment not normal.