Charles Stein: Semiparametric Models and Nonparametric Methods

Peter Bickel

UC Berkeley

March, 2010

(With assistance of Jing Lei)
Outline


Discussion: I will sketch Stein’s contributions primarily in I and II, and indicate how ideas have been applied and developed.
Charles Stein
• I: A fundamental paper whose importance was not fully appreciated till 1970’s - 1980’s.

• The basic question: Given $X_1, \ldots, X_n$ iid from $p(x, \theta, \gamma, \eta)$, $\theta \in \mathbb{R}^p$, $\gamma \in \mathbb{R}^q$, $\eta \in H$ infinite dimensional.
  
  • When can one make inference about $\theta$ as well not knowing $\eta$ as knowing it?
  
  • Stein used both a testing framework and estimation.
  
  • will use estimation only.
The basic idea

- The parametric case as a guide to semiparametric. Suppose \( \eta \) is \( \ell \) dimensional.

- There typically exists at each \((\theta_0, \gamma_0, \eta_0) \equiv \theta_0\) a “least favorable” \( p \) dimensional parametric sub-family \( p(\theta_0 + s, \gamma_0 + t(s), \eta_0 + \lambda(s))\), where \( t(\cdot) \) and \( \lambda(\cdot) \) depend on \((\theta_0, \gamma_0, \eta_0)\), in which estimation of \( \theta \) is as hard as for the original model.

**NB** All conclusions are asymptotic, all models are smoothly parametrized.
Example:

\[ Z_{i1}, Z_{i2} \text{ iid } N(0,1), \ p = q = \ell = 1. \]

\[ X_{i1} = \gamma + \eta Z_{i1} \]
\[ X_{i2} = \gamma + \theta + \eta Z_{i2} \]
\[ \hat{\theta} = \bar{X}_2 - \bar{X}_1. \]

\[ \bullet \text{ At } \theta_0 = \gamma_0 = 0, \ \sigma = 1 \]
\[ \gamma(\theta) = -\frac{\theta}{2} \]
The principle holds for $l = \infty$

3 examples:

i) Symmetric location.

ii) 2-sample location and scale.

iii) Signal in Gaussian noise.
Basic heuristic tool

If

\[
\Sigma = \begin{pmatrix}
\Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\
\Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\
\Sigma_{31} & \Sigma_{32} & \Sigma_{33}
\end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix}
\Sigma^{11} & \Sigma^{12} & \Sigma^{13} \\
\Sigma^{21} & \Sigma^{22} & \Sigma^{23} \\
\Sigma^{31} & \Sigma^{32} & \Sigma^{33}
\end{pmatrix},
\]

\(\Sigma_{11} \in \mathbb{R}^{p \times p}, \Sigma_{22} \in \mathbb{R}^{q \times q}, \Sigma_{33} \in \mathbb{R}^{\ell \times \ell}\).

\[
\begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix}^{-1} = \begin{pmatrix}
\Sigma^{11} & \Sigma^{12} \\
\Sigma^{21} & \Sigma^{22}
\end{pmatrix} \quad \text{iff} \quad \Sigma_{13} = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{23}
\]

\((*)\)
Meaning

- If $\Sigma = I(\theta_0)$ Fisher information, $\theta$ is no harder to estimate in presence of unknown $\gamma$ and $\eta$ at $(\theta_0, \gamma_0, \eta_0)$ than just in presence of unknown $\gamma$ iff (*) holds.

- Even if $\eta$ is $\infty$ dimensional, a $p$-dimensional least favorable model $p(\theta_0 + t, \gamma_0 + \gamma(t), \eta_0)$ will be as hard for estimating $\theta$ as in full model $p(\theta_0, \gamma, \eta)$.

**NB** The family depends (in general) on where you are in the model.
Symmetric location

\[ p(x, \theta, g) = g(x - \theta), \ g \text{ symmetric, } \int \frac{[g']^2}{g} < \infty. \]

\( q = 0, \ \text{WLOG } \theta_0 = 0. \)

**Condition:** \( l_{12} = 0, \int \frac{g_0'(x)}{g_0(x)} \frac{\partial}{\partial \eta} \log g_\eta(x) \bigg|_{\eta=0} g_0(x) \ dx = 0 \) at \((\theta_0, g_0).\)

\( g_0(x - \theta) \)

Implication: possibility of constructing estimate using only data which does “as well as MLE” for each shape.
(iii) Linear relation/Signal in Gaussian noise.

- $X_i = Y_i + Z_i$
- $X_i = (X_{i1}, X_{i2})$ etc.
- $Y_i$ non-Gaussian, $P[Y_i \in L] = 1$, distribution unknown.
- $Z_i$ bivariate Gaussian, $Y_i \perp Z_i$.
- It should be no harder to estimate the slope of $L$ knowing the distributions of $Y$, $Z$ up to affine transformation as not knowing them.
Developments in ’70’s, ’80’s, ’90’s

I. Many surprising models

- Cox model (biostatistics)
- Censoring/truncation in regression etc. (biostatistics, engineering, econometrics, astronomy)
- Signal in Gaussian noise, Independent Component Analysis (engineering).
- and many more...
II. Theoretical results

- Geometrical approach to finding least favorable families.
- Relations to robust estimation and work of Le Cam.
- Carry over into minimax analysis in nonparametric statistics. Donoho, Liu and others.
II: an inequality with ever developing uses. One of elements of revival of an important expansion.

The inequality:

Given $X_1, \ldots, X_n$ iid, $T_k \equiv (X_1, \ldots, X_k)$ symmetric,

$1 \leq k \leq n - 1$.

$T_k \in L_2(P)$, all $k$.

$T_n^{(-i)} \equiv T_{n-1}(X_j : j \neq i)$. 
Jackknife pseudo value

\[ \tau_i \equiv nT_n - (n - 1)T_n^{(-i)} \]
\[ \widetilde{\text{Var}}_n \equiv \sum_{i=1}^{n} (\tau_i - \bar{\tau})^2 \]

E-S inequality:

\[ \text{Var} T_{n-1} \leq E(\widetilde{\text{Var}} T_n). \]
The Hoeffding/ANOVA expansion

For $X_1, \ldots, X_n$ independent, $T_n$ general.

If $T_n \in L_2(P)$,

$$T_n = \mu + \sum_i A_i(X_i) + \sum_{i<j} A_{ij}(X_i, X_j) + \ldots,$$

where all terms in the expansion are orthogonal in $L_2(P)$. 
A form for terms

\[ T_n(X_1, \ldots, X_n) = \mu + \int T_n(y_1, \ldots, y_n) \left( \prod_{i=1}^{n} \delta x_i(dy_i) - \prod_{i=1}^{n} dP_i(y_i) \right) \]

\( \mu = EP_T. \)

Write out Taylor expansion of difference. Terms are

\[ A_S(X_i : i \in S) = \int \ldots \int T_n(y_1, \ldots, y_n) \prod_{i \in S} (\delta x_i - P)(dy_i) \]

\[ \times \int T_n(y_1, \ldots, y_n) \prod_{i \in \bar{S}} dP(y_i) \]

where \( S \) ranges over all subsets of \( \{1, \ldots, n\} \).
Major terms

\[ \sum \{ A_S(X_i : i \in S) : |S| = m \} = \Pi(T|W_m), \]  
where \( V_1 \subset \cdots \subset V_n \) are subspaces of \( L_2(P) \) corresponding to sums of functions of single \( X_i \)'s, pair of \( X_i \)'s, etc.

\[ W_j = V_j \cap V_{j-1}', \ \Pi \leftrightarrow \text{projection in } L_2(P). \]

If \( X_i \) are iid, \( T_n \) symmetric, then \( A_S \equiv A_{|S|}. \)
\[
\widetilde{\text{Var}}(T_n) = \frac{1}{n} \sum_{i < i'} E \left( T_n^{(-i)} - T_n^{(-i')} \right)^2.
\]

Proof of inequality

- Write out \( T_{n-1}(X_j : j \neq i), \ T_n^{(-i')}, \ T_n \) terms of expansion using \( A_S \equiv A_{|S|} \).

- Compute squared norm using orthogonality and
  \[
  EA_S^2(X_j : j \in S) = EA_m^2(X_1, \ldots, X_m) \text{ if } |S| = m,
  \]

- Compare coefficients.
Immediate related and independent results

Karlin-Rinott (1982): E-S ⇐

\[ \| E(T_n|X_1, \ldots, X_{r+1}) - E(T_n|X_1, \ldots, X_r) \|^2 \gtrsim \text{ in } r \leq n - 1. \]

Show more generally,

\[ (-1)^{s+1} D^s \| E(T|X_1, \ldots, X_r) \|^2 \geq 0 \text{ for all } s \geq 1 \quad (\ast) \]
Immediate related and independent results

van Zwet (1983):

• \((\ast)\) (independently derived) + Fourier theory.

• General Berry-Esseen Theorem for \(T_n\) symmetric.

• Since then Efron-Stein has
  • 200 citations dealing with
  • Applications in mathematics, econometrics, finance, medical informatics,...
B. Eckmann’s Dictum *

Thou shalt not publish more than once a year

(But when you do make it count)

Charles has (without knowing it) excelled according to that dictum throughout his career.

Happy Birthday, Charles!

*Algebraic topologist at ETH, Peter Huber’s thesis advisor.