

# Efficient Functional Estimation in Semiparametric Models

**Yanyuan Ma**

Université de Neuchâtel  
Texas A&M University  
Yanyuan.Ma@unine.ch

<http://www.stat.tamu.edu/~ma>

Joint work with Arnab Maity and Raymond Carroll.

## Outline

1. Motivation from EATS data
2. Literature review and empirical discovery
3. Main results
4. EATS data analysis
5. Summary

## EATS Data: Summary

- Eating at America's Table Study (EATS) (Subar, et al. 2001)
- Interest: intake of various dietary components (e.g. red meat)
- Usual intake: average daily intake by an individual
- Goal: estimation of the distribution of usual intake

## EATS Data: Specifics

- $n = 886$  individuals
- Data consist of four 24hr recalls over the course of a year
- Response: intake in ounces ( $Y$ )
- Covariates measured are
  - $X = (\text{age, gender})$
  - $Z = \log(\text{energy in calories})$ .
- 45% of the 24-hour recalls reported no red meat consumption

**EATS Data: Model**

- Logit model for red meat consumption indicator

$$P_r(\Delta_{ij} = 1 | X_i, Z_i, U_{i1}) = H(\beta_0 + X_i^T \beta_1 + U_{i1})$$

- Lognormal model for nonzero responses

$$[\Delta_{ij} \log(Y_{ij}) | \Delta_{ij} = 1, X_i, Z_i, U_{i2}] \sim \text{Normal}\{X_i^T \beta_2 + \theta(Z_i) + U_{i2}, \sigma^2\}$$

- Usual intake for an individual is defined as

$$\begin{aligned} & G\{X, U_1, U_2, \beta, \theta(Z)\} \\ &= \exp\{X^T \beta_2 + \theta(Z) + U_2 + \sigma^2/2\} H(\beta_0 + X^T \beta_1 + U_1) \end{aligned}$$

- Goal:  $\kappa = P_r[G\{X, U_1, U_2, \beta, \theta(Z)\} > c]$  (or  $E(G)$ ).

**EATS Data: Simplification**

- Define

$$\begin{aligned}\mathcal{F}\{X, \beta, \theta(Z)\} &= P_r[G\{X, U_1, U_2, \beta, \theta(Z)\} > c | X, Z] \\ \kappa &= E[\mathcal{F}\{X, \beta, \theta(Z)\}]\end{aligned}$$

- Define

$$\begin{aligned}\mathcal{F}\{X, \beta, \theta(Z)\} &= P_r[G\{X, U_1, U_2, \beta, \theta(Z)\} | X, Z] \\ \kappa &= E[\mathcal{F}\{X, \beta, \theta(Z)\}]\end{aligned}$$

- Common problem: estimate a functional of  $\beta$  and  $\theta$ .

## General Problem

- Model: loglikelihood  $l\{Y|X, Z; \beta, \theta(Z)\}$
- True parameter:  $\beta_0, \theta_0(Z)$
- Conditions:  
 $l$  satisfies

$$E[\partial l\{Y, X, \theta_0(Z), \beta_0\} / \partial \beta | X, Z] = 0$$

$$E[\partial l\{Y, X, \theta_0(Z), \beta_0\} / \partial \{\theta(Z)\} | X, Z] = 0$$

$l$  admits a finite information matrix.

$\mathcal{F}$  differentiable

- Goal: Estimate  $\kappa = E[\mathcal{F}\{X, \theta(Z), \beta\}]$ .

## Literature Review

- Obvious estimator:  $\hat{\kappa} = n^{-1} \sum_{i=1}^n \mathcal{F}(X_i, \hat{\theta}(Z_i, \hat{\beta}), \hat{\beta})$ .
- Naive estimators in partial linear models:

- $\kappa = E(Y) :$

$$\hat{\kappa} = n^{-1} \sum_{i=1}^n Y_i \quad \hat{\kappa} = n^{-1} \sum_{i=1}^n X_i^T \hat{\beta} + \hat{\theta}(Z_i, \hat{\beta})$$

Both are efficient (Wang et al. 2004).

- $\kappa = P_r(Y > c) :$

$$\hat{\kappa} = n^{-1} \sum_{i=1}^n I(Y_i > c) \quad \hat{\kappa} = n^{-1} \sum_{i=1}^n I\{X_i^T \hat{\beta} + \hat{\theta}(Z_i, \hat{\beta}) > c\}$$

One is often much more efficient than the other.

**Main Result I**

$\hat{\kappa} = n^{-1} \sum_{i=1}^n \mathcal{F}(X_i, \hat{\theta}(Z_i, \hat{\beta}), \hat{\beta})$  is semiparametric efficient.

If

- $\theta$  estimated by kernel methods
- $h = o_p(n^{-1/4})$
- $\beta$  estimated by profiling or backfitting
- $\kappa$  is pathwise differentiable

**Main Result II**

For exponential family

$$f(y|x, z) = \exp \left[ \frac{yc\{\eta(x, z)\} - \mathcal{C}[c\{\eta(x, z)\}]}{\phi} + \mathcal{D}(y, \phi) \right],$$
$$\eta(x, z) = x^T \beta + \theta(z)$$

Sample mean is efficient iff

- $\partial c\{x^T \beta + \theta(Z)\} / \partial \theta(Z)$  is a function of  $Z$  only (canonical).
- $c\{x^T \beta + \theta(z)\} = a + b \log\{x^T \beta + \theta(z)\}$ .

Special case:

$$Y = X^T \beta + \theta(Z) + \epsilon \quad \epsilon \sim N(0, \sigma^2)$$

**Main Result III**

- The  $\pi^{\text{th}}$  quantile ( $q$ ):  $\mathcal{F}\{q, X, \theta(Z), \beta\} = \text{pr}(Y \leq q|X, Z)$
- $E[\mathcal{F}\{q, X, \theta(Z), \beta\}] = \pi$
- Obvious estimator: solve for  $\hat{q}$  from
$$n^{-1} \sum_{i=1}^n \mathcal{F}\{\hat{q}, X_i, \hat{\theta}(Z_i, \hat{\beta}), \hat{\beta}\} - \pi = 0$$
- $\hat{q}$  is semiparametric efficient if  $\mathcal{F}(\bullet)$  is strictly monotone in an open neighborhood of  $q_0$  and differentiable at  $q_0$ .

**Main Result IV**

Single index model:  $\theta(Z) \longrightarrow \theta(Z^T \alpha)$

$\hat{\kappa} = n^{-1} \sum_{i=1}^n \mathcal{F}\{X_i, \hat{\theta}(Z_i^T \hat{\alpha}, \hat{\beta}), \hat{\beta}\}$  is semiparametric efficient.

If

- $\theta$  estimated by kernel methods
- $h = o_p(n^{-1/4})$
- $\beta, \alpha$  estimated by profiling or backfitting
- $\kappa$  is pathwise differentiable
- other regularity conditions (Carroll et al, 1997)

**EATS Data: Analysis**

Naive estimator is inconsistent!

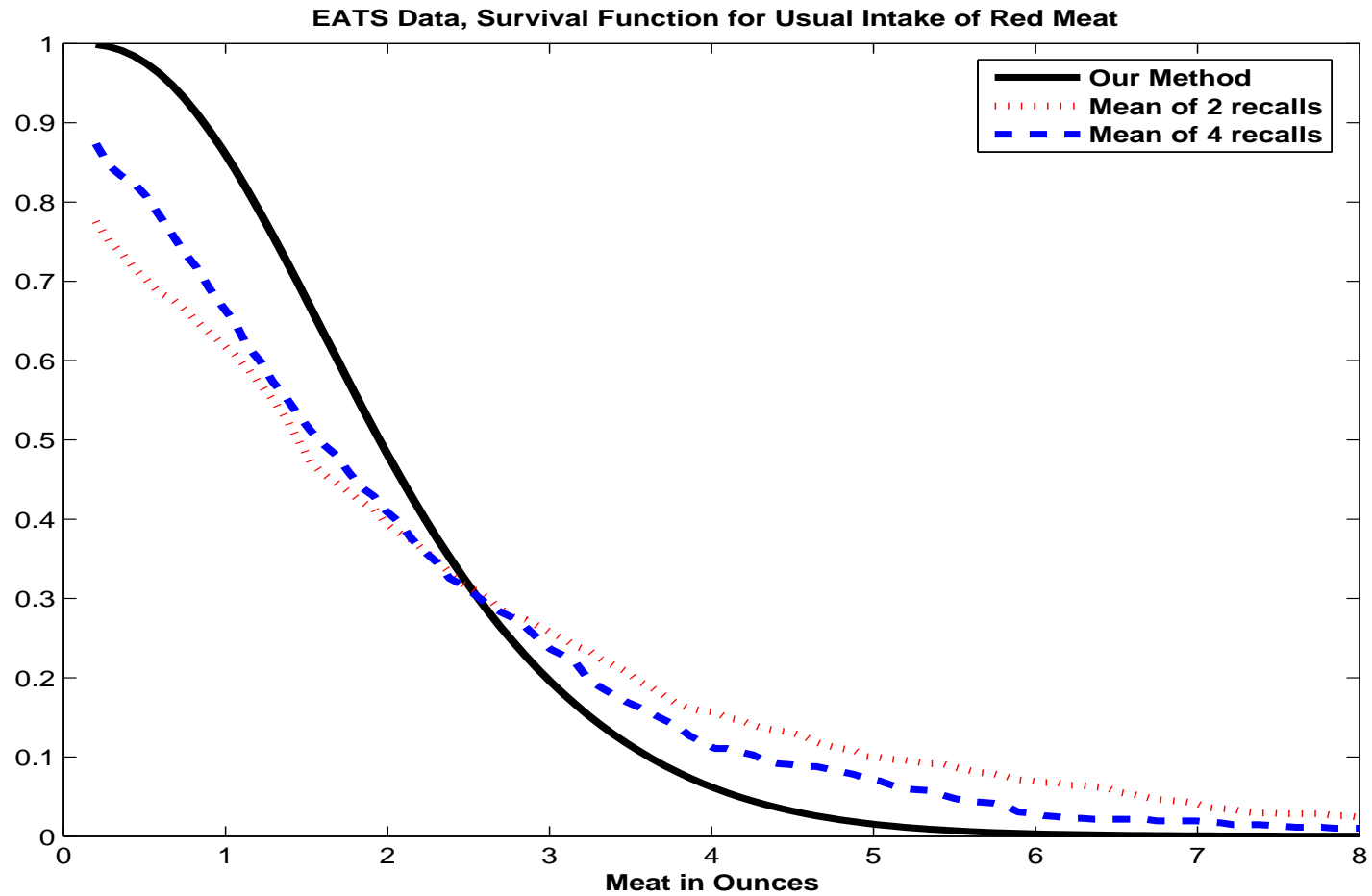
Suppose only one 24-hour recall were computed and the percentage of these 24-hour recalls exceeding  $c$  is computed. In large samples, this percentage converges to

$$\begin{aligned} & \kappa_{24\text{hr}} \\ &= E \left\{ \Phi \left( \left[ X^T \beta_2 + \theta(Z) - \log\{c/H(\beta_0 + X^T \beta_1 + U_1)\} \right] / (\sigma^2 + \sigma_2^2)^{1/2} \right) \right\} \end{aligned}$$

In contrast,

$$\kappa_0 = E \left\{ \Phi \left( \left[ X^T \beta_2 + \theta(Z) + \sigma^2/2 - \log\{c/H(\beta_0 + X^T \beta_1 + U_1)\} \right] / \sigma_2 \right) \right\}$$

## EATS Data: Result



## Summary

- Functional estimation in semiparametric models
- Naive estimator can be inconsistent or inefficient
- Obvious plug-in estimator is efficient
- **Sometimes**, naive estimator could be efficient and robust
- Efficiency theory relies on a different kind of geometry