Generalized Thresholding Estimators for High Dimensional Location Parameters

Min Zhang

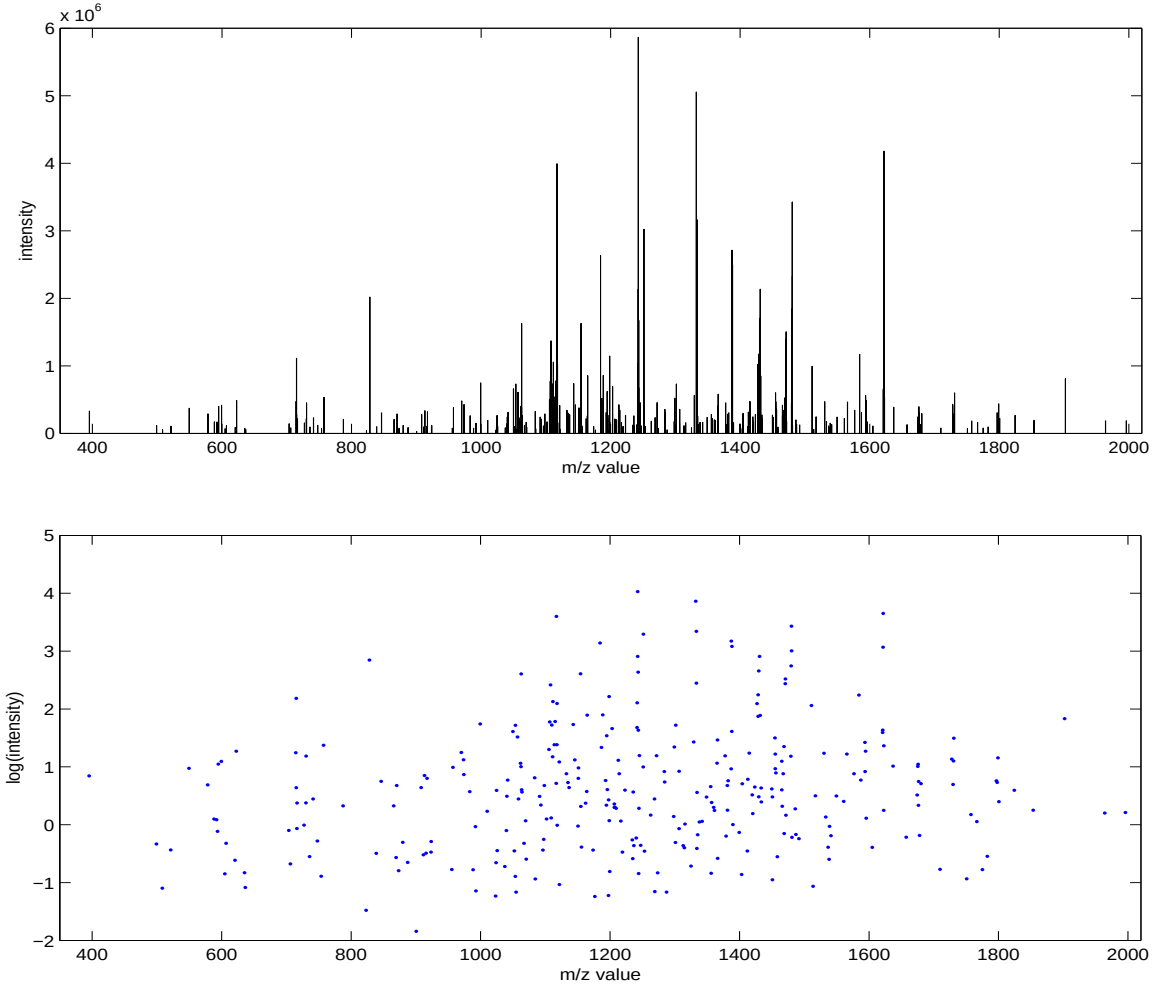
Department of Statistics

Purdue University

Joint work with Dabao Zhang and Martin Wells

May 16, 2009

Mass Spectrometry Data



Original data (top) and the normalized peaks (bottom).

From Keller et al., (2002)

Introduction

- Data:

$$\mathbf{Y_p} = (y_1, y_2, \ldots, y_p)$$

- Model:

$$y_i - \mu_i \overset{iid}{\sim} \varphi(\cdot),$$

where $\varphi(\cdot)$ is a symmetric log-concave density function.

- Goal:

Estimate the high dimensional location parameters $\mu = (\mu_1, \mu_2, \ldots, \mu_p)$ from the noisy data.

Generalized Thresholding Estimators

- Definition: *For* $\tau_- \leq 0$ *and* $\tau_+ \geq 0$, $\delta(y, \tau_-, \tau_+)$ *is a generalized thresholding estimator if (i)* $\delta(y, \tau_-, \tau_+)$ *is increasing on* $y \in \mathbb{R}$*; (ii)* $-|y| \leq \delta(y, \tau_-, \tau_+) \leq |y|, \forall y \in \mathbb{R}$*; (iii)* $\delta(y, \tau) = \delta(y, -\tau, \tau)$ *is antisymmetric for any* $\tau \geq 0$*; (iv)* $\delta(y, \tau_-, \tau_+) = 0$ *if and only if* $\tau_- \leq y \leq \tau_+$. This is a generalization of the empirical Bayes estimator proposed by Johnstone and Silverman (2004)
- A Bayes Construction:
  $\mu_i \overset{iid}{\sim} (1 - w_- - w_+)\delta_0(\mu) + w_-\gamma_-(\mu) + w_+\gamma_+(\mu)$,
  and $\hat{\mu}(y_i; w_-, w_+) = \text{median}(\mu_i | y_i; w_-, w_+)$.
- Prior:
$$\begin{cases} \gamma_+(\mu) = \sqrt{\frac{2}{\pi}} \left(1 - \frac{\mu(1-\Phi(\mu))}{\phi(\mu)}\right) 1_{[0,\infty)}(\mu) \\ \gamma_-(\mu) = \sqrt{\frac{2}{\pi}} \left(1 + \frac{\mu\Phi(\mu)}{\phi(\mu)}\right) 1_{(-\infty,0]}(\mu) \end{cases}$$

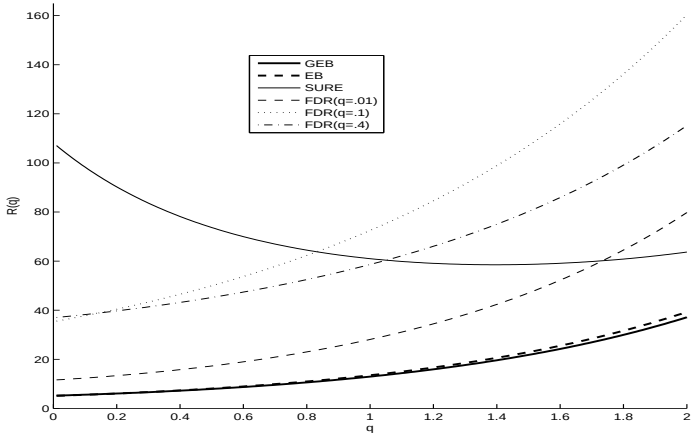
Simulation Study

Comparing the GEB estimator with others when $k_-/k_+ = 5/45$ and $|\mu_-| \neq \mu_+$. $R(2)$, NFP, and NFN are recorded for each estimator with the largest values in bold type. The corresponding standard errors are bracketed.

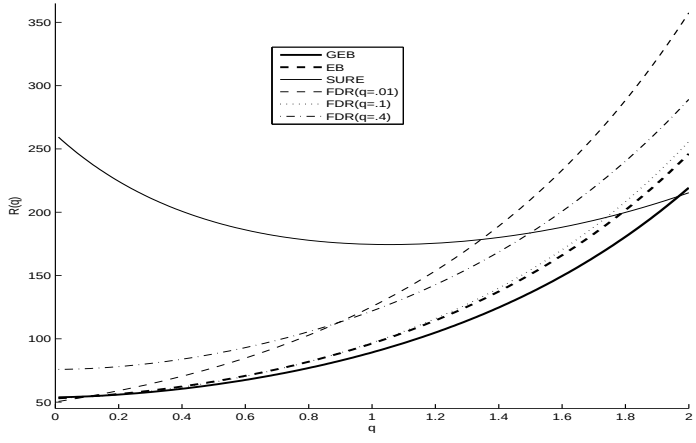
μ_-/μ_+	Criterion	GEB	EB	SURE	FDR ($q = .01$)	FDR ($q = .1$)
$-5/3$	$R(2)$	219 (1.15)	246 (1.05)	215 (0.88)	**358** (1.06)	256 (1.12)
	NFP	3.74 (0.07)	2.95 (0.06)	**212** (2.18)	0.15 (0.01)	3.11 (0.06)
	NFN	17.71 (0.13)	21.99 (0.13)	1.96 (0.06)	**35.85** (0.12)	21.62 (0.13)
$-7/3$	$R(2)$	212 (1.01)	241 (1.00)	215 (0.88)	**343** (0.84)	253 (1.07)
	NFP	4.35 (0.11)	2.99 (0.06)	**212** (2.18)	0.15 (0.01)	3.12 (0.06)
	NFN	17.29 (0.13)	21.78 (0.13)	1.96 (0.06)	**35.09** (0.11)	21.48 (0.13)
$-7/5$	$R(2)$	92.62 (0.70)	99.83 (0.79)	221 (0.90)	118 (1.36)	111 (0.98)
	NFP	7.25 (0.10)	6.47 (0.09)	**220** (1.99)	0.47 (0.02)	5.28 (0.08)
	NFN	0.24 (0.01)	0.47 (0.02)	0.01 (0.00)	**2.93** (0.05)	0.58 (0.02)

Simulation Study

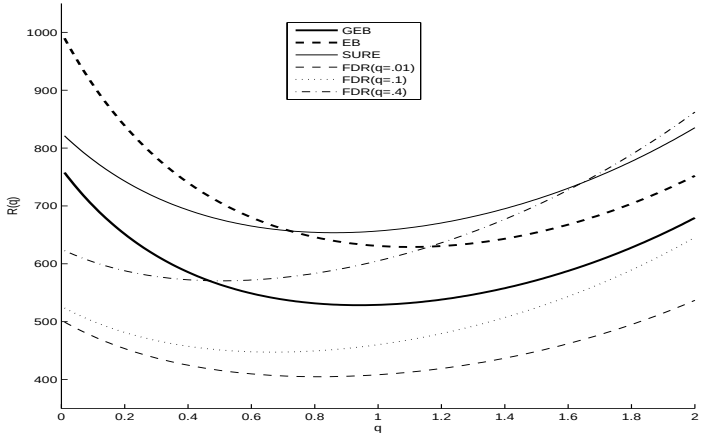
a. $(\mu_-, \mu_+) = (-3, 3)$, $(k_-, k_+) = (0, 5)$



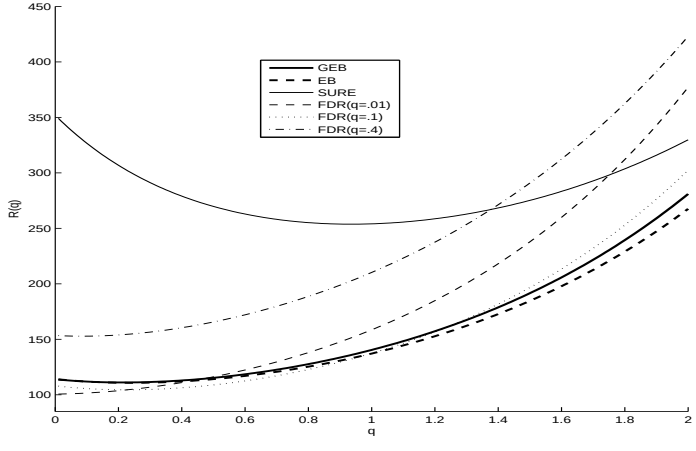
b. $(\mu_-, \mu_+) = (-5, 3)$, $(k_-, k_+) = (5, 45)$



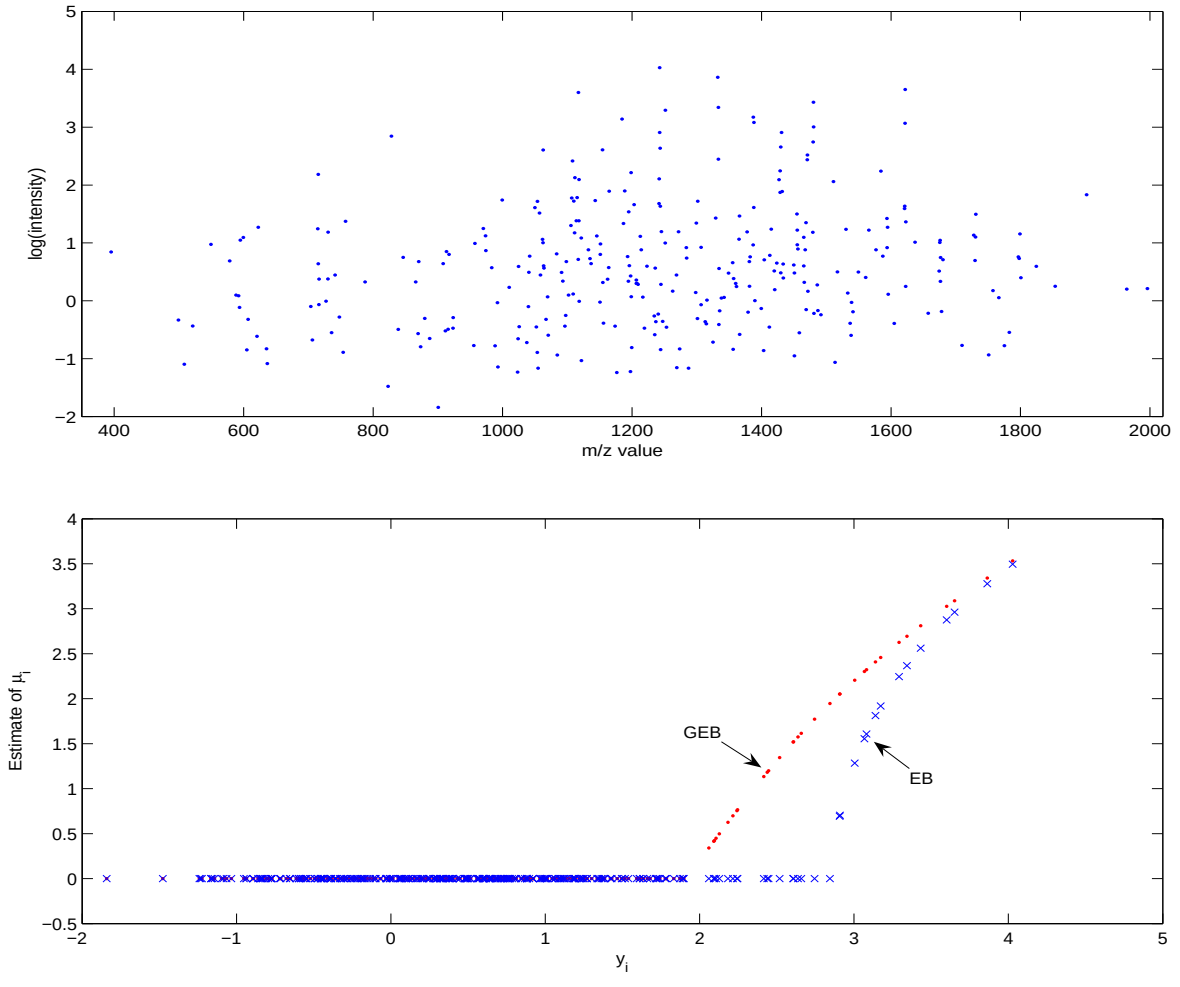
c. $(\mu_-, \mu_+) = (-7, 5)$, $(k_-, k_+) = (450, 50)$



d. $(\mu_-, \mu_+) = (-7, 3)$, $(k_-, k_+) = (50, 50)$



Mass Spectrometry Data Analysis



Normalized peaks of the original data (top) and identified peaks (bottom).

THANK YOU