Emprical Bayes Estimators for Sparse Sequences

Pavan Srinath        Ramji Venkataramanan

University of Cambridge

ISIT 2018
Sparse Vector Estimation

Vector $\theta \in \mathbb{R}^n$ to be estimated from observation $y = \theta + w$

$$
\begin{bmatrix}
y_1 \\
\vdots \\
y_n
\end{bmatrix} =
\begin{bmatrix}
\theta_1 \\
\vdots \\
\theta_n
\end{bmatrix} +
\begin{bmatrix}
w_1 \\
\vdots \\
w_n
\end{bmatrix}, \quad w_i \text{ i.i.d. } \sim \mathcal{N}(0, 1)
$$

- A fraction $\eta$ of the entries are non-zero
- Sparsity level $\eta$ may be unknown to the estimator
Applications

\[ y = \theta + w \]

Function estimation

Estimate function \( f \) from \( y(t) = f(t) + w(t), \quad t \in [0, T] \)

\( f \) has sparse representation in suitable orthonormal basis

Estimate basis coefficients \( \theta \) from noisy observation \( y \)
Applications

\[ y = \theta + w \]

Function estimation

Estimate function \( f \) from \( y(t) = f(t) + w(t), \ t \in [0, T] \)

\( f \) has sparse representation in suitable orthonormal basis

Estimate basis coefficients \( \theta \) from noisy observation \( y \)

Compressed sensing

Wish to recover \( \theta \) from linear measurements \( A\theta + \text{noise} \)

AMP: iterative algorithm that generates estimates \( \hat{\theta}_1, \hat{\theta}_2, \ldots \)

\( \hat{\theta}_{t+1} \) generated from \( y_t \approx \theta + \tau_t w \) using a sparse estimator
Performance

\[ y = \theta + w \]

Loss function of estimator \( \hat{\theta}(y) \) is \( \| \theta - \hat{\theta}(y) \|^2 \)

The *normalized risk* of the estimator is

\[
R(\theta, \hat{\theta}) = \frac{1}{n} \mathbb{E} \left[ \| \hat{\theta}(y) - \theta \|^2 \right]
\]

Expectation is calculated with the density \( y \sim N(\theta, I) \)
Thresholding Estimators

\[ y = \theta + w \]

**Hard Thresholding**

Fix threshold \( \lambda > 0 \), set entries with \( |y_i| < \lambda \) to zero
Thresholding Estimators

\[ y = \theta + w \]

**Soft Thresholding:** Also shrink other entries towards zero by \( \lambda \)

\[ \hat{\theta}_{ST}(y; \lambda) = \begin{cases} 
  y - \lambda, & y \geq \lambda \\
  0, & -\lambda < y < \lambda \\
  y + \lambda, & y \leq -\lambda
\end{cases} \]

- For large \( n \) and sparsity level \( \eta \to 0 \), worst-case risk over set of \( \eta \)-sparse vectors is \( 2\eta \log \eta^{-1}(1 + o(1)) \)
- Close to minimax optimal

Can we do better, particularly for moderate or large values of \( \eta \)?

---

James-Stein estimators

\[ y = \theta + w \]

First consider case where \( \theta \) is not (necessarily) sparse

- **ML estimator:** \( \hat{\theta}_{\text{ML}} = y \)

- **Shrinkage estimator (James-Stein ’61):**

\[
\hat{\theta}_{\text{JS}} = \left( 1 - \frac{n - 2}{\|y\|^2} \right) y
\]

Dominates \( \hat{\theta}_{\text{ML}} \) for all \( \theta \)
James-Stein estimators

\[ y = \theta + w \]

First consider case where \( \theta \) is not (necessarily) sparse

- ML estimator: \( \hat{\theta}_{\text{ML}} = y \)

- Shrinkage estimator (James-Stein '61):
  \[
  \hat{\theta}_{\text{JS}} = \left( 1 - \frac{n - 2}{\|y\|^2} \right)_+ y
  \]
  Dominates \( \hat{\theta}_{\text{ML}} \) for all \( \theta \)

- Shrinkage estimator (Lindley '62):
  \[
  \hat{\theta}_{\text{L}} = \bar{y}1 + \left( 1 - \frac{n - 3}{\|y - \bar{y}1\|^2} \right)_+ (y - \bar{y}1), \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i
  \]
  Also dominates \( \hat{\theta}_{\text{ML}} \) for all \( \theta \)
Empirical Bayes

\[ y = \theta + w \]

If we assume \( \theta_i \) i.i.d. \( \sim N(\mu, \xi^2) \), then

\[ \hat{\theta}_{\text{Bayes}} = \mu 1 + \left( 1 - \frac{1}{1 + \xi^2} \right) (y - \mu 1) y \]

Empirical Bayes

\[ y = \theta + w \]

If we assume \( \theta_i \) i.i.d. \( \sim N(\mu, \xi^2) \), then

\[ \hat{\theta}_{\text{Bayes}} = \mu 1 + \left(1 - \frac{1}{1 + \xi^2}\right)(y - \mu 1) y \]

Plug-in estimates for \( \mu \) and \( 1/(1 + \xi^2) \), based on

\[ \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} y_i \right] = \mu \quad \text{and} \quad \mathbb{E} \left[ \frac{n - 3}{\| y - \mu 1 \|^2} \right] = \frac{1}{1 + \xi^2}, \]

This gives \( \hat{\theta}_L = \bar{y} 1 + \left(1 - \frac{n-3}{\| y - \bar{y} 1 \|^2}\right)(y - \bar{y} 1) \)

Empirical Bayes for sparse $\theta$

Sparsity-promoting prior for each entry of $\theta$:

$$f(\theta; \varepsilon, \mu, \xi) = (1 - \varepsilon)\delta(\theta) + \varepsilon \psi(\theta; \mu, \xi), \quad \theta \in \mathbb{R}.$$  

- Mixture of point mass at 0 and continuous distribution $\psi$
- Parameter $\varepsilon \in (0, 1)$ controls sparsity — need not be equal to true sparsity $\eta$
- $\mu$ is location parameter (mean), $\xi$ is scale parameter
In this work

\[ f(\theta; \varepsilon, \mu, \xi) = (1 - \varepsilon)\delta(\theta) + \varepsilon \psi(\theta; \mu, \xi), \quad \theta \in \mathbb{R}. \]

Take \( \psi \) to be Gaussian with mean \( \mu \), variance \( \xi^2 \)

Empirical Bayes estimator \( \hat{\theta}_{EB} \) that estimates \( \hat{\mu}, \hat{\xi}^2 \) from data
\( \varepsilon \) treated as fixed parameter that can be optimized
In this work

\[ f(\theta; \varepsilon, \mu, \xi) = (1 - \varepsilon)\delta(\theta) + \varepsilon \psi(\theta; \mu, \xi), \quad \theta \in \mathbb{R}. \]

Take \( \psi \) to be Gaussian with mean \( \mu \), variance \( \xi^2 \)

Empirical Bayes estimator \( \hat{\theta}_{EB} \) that estimates \( \hat{\mu}, \hat{\xi}^2 \) from data

\( \varepsilon \) treated as fixed parameter that can be optimized

- Analyze loss, risk of \( \hat{\theta}_{EB} \)
- Hybrid estimator that reliably picks better of \( \hat{\theta}_{EB} \) and \( \hat{\theta}_{ST} \)
The Empirical Bayes Estimator

\[ y = \theta + w \]

\[ f(\theta; \varepsilon, \mu, \xi) = (1 - \varepsilon) \delta(\theta) + \varepsilon \psi(\theta; \mu, \xi), \quad \theta \in \mathbb{R}. \]

\[ \hat{\mu}(y) = \bar{y}, \quad \hat{\xi}^2(y) = \frac{1}{\varepsilon} \left( \frac{\bar{y}^2 - (\bar{y})^2}{\varepsilon} - 1 \right) + \]
The Empirical Bayes Estimator

\[ y = \theta + w \]

\[ f(\theta; \varepsilon, \mu, \xi) = (1 - \varepsilon) \delta(\theta) + \varepsilon \psi(\theta; \mu, \xi), \quad \theta \in \mathbb{R}. \]

\[ \hat{\mu}(y) = \frac{\bar{y}}{\varepsilon}, \quad \hat{\xi}^2(y) = \frac{1}{\varepsilon} \left( \frac{y^2}{\varepsilon} - \frac{(\bar{y})^2}{\varepsilon} - 1 \right) \]

\[ \hat{\theta}_{EB,i}(y; \varepsilon) = \frac{\hat{\mu} + \left( 1 - \frac{1}{1 + \hat{\xi}^2} \right) (y_i - \hat{\mu})}{1 + \left( \frac{1 - \varepsilon}{\varepsilon} \right) \sqrt{1 + \hat{\xi}^2} \exp \left( -\frac{y_i^2}{2} + \frac{(y_i - \hat{\mu})^2}{2(1 + \hat{\xi}^2)} \right)}, \quad i \in [n]. \]
The Empirical Bayes Estimator

\[ y = \theta + w \]

\[ f(\theta; \varepsilon, \mu, \xi) = (1 - \varepsilon) \delta(\theta) + \varepsilon \psi(\theta; \mu, \xi), \quad \theta \in \mathbb{R}. \]

\[ \hat{\mu}(y) = \frac{\bar{y}}{\varepsilon}, \quad \hat{\xi}^2(y) = \frac{1}{\varepsilon} \left( \frac{y^2}{\varepsilon} - \left( \frac{\bar{y}}{\varepsilon} \right)^2 - 1 \right) + \]

For simplicity, we will assume \( \hat{\mu} = 0 \). Then

\[ \hat{\theta}_{EB,i}(y; \varepsilon) = \frac{\left( 1 - \frac{1}{1 + \hat{\xi}^2} \right) y_i}{1 + \left( 1 - \varepsilon \right) \sqrt{1 + \hat{\xi}^2} \exp \left( - \frac{\hat{\xi}^2 y_i^2}{2(1 + \hat{\xi}^2)} \right)} \]
Empirical Bayes vs Soft-thresholding

\[ \tilde{R}(\theta, \hat{\theta}) / n \]

\( \theta \) with all non-zero \( \theta_i = 3 \)
Empirical Bayes vs Soft-thresholding

\[ \tilde{R}(\theta, \hat{\theta})/n \]

\( \theta \) with non-zero \( \theta_i \in \{3, -3\} \)
Risk Estimate

\[ y = \theta + w \]

Stein’s Unbiased Risk Estimate (SURE) [Stein ’81]

For an estimator \( \hat{\theta}(y) \) is differentiable almost everywhere,

\[
\hat{R}(\theta, \hat{\theta}(y)) := -1 + \frac{1}{n} \| y - \hat{\theta} \|^2 + \frac{2}{n} \sum_{i=1}^{n} \frac{\partial \hat{\theta}_i}{\partial y_i},
\]

is an unbiased estimate of the normalized risk \( R(\theta, \hat{\theta}) \), i.e.,

\[
\mathbb{E}[\hat{R}(\theta, \hat{\theta}(y))] = R(\theta, \hat{\theta})
\]
Risk Estimate

\[ y = \theta + w \]

Stein's Unbiased Risk Estimate (SURE) [Stein '81]

For an estimator \( \hat{\theta}(y) \) is differentiable almost everywhere,

\[
\hat{R}(\theta, \hat{\theta}(y)) := -1 + \frac{1}{n} \| y - \hat{\theta} \|^2 + \frac{2}{n} \sum_{i=1}^{n} \frac{\partial \hat{\theta}_i}{\partial y_i},
\]

is an unbiased estimate of the normalized risk \( R(\theta, \hat{\theta}) \), i.e.,

\[
\mathbb{E}[\hat{R}(\theta, \hat{\theta}(y))] = R(\theta, \hat{\theta})
\]

The SURE for soft-thresholding with threshold \( \lambda \) is

\[
\hat{R}(\theta, \hat{\theta}_{ST}; \lambda) = -1 + \frac{\| y - \hat{\theta}_{ST} \|^2}{n} + \frac{2}{n} \sum_{i=1}^{n} 1\{y_i^2 > \lambda^2\}.\]
SURE for Empirical Bayes estimator

\[ \hat{R}(\theta, \hat{\theta}_{EB}; \varepsilon) = \left( \frac{\|y\|^2}{n} - 1 \right) + \frac{a_y^2}{n} \sum_{i=1}^{n} y_i^2 (1 + 2 cy e^{-\frac{a_y y_i^2}{2}}) \]

\[ \quad - \frac{2 a_y}{n} \sum_{i=1}^{n} \frac{y_i^2 - 1}{b_{y,i}} + O \left( \frac{1}{n} \right) \]

where

\[ a_y := \frac{\hat{\xi}^2}{1 + \hat{\xi}^2}, \quad c_y := \frac{1 - \varepsilon}{\varepsilon} \sqrt{1 + \hat{\xi}^2}, \]

\[ b_{y,i} := 1 + c_y e^{-\frac{a_y y_i^2}{2}}. \]
Hybrid Estimator

Choose estimator by comparing risk estimates:

\[ \hat{\theta}_H = \begin{cases} 
\hat{\theta}_{EB} & \text{if } \hat{R}(\theta, \hat{\theta}_{EB}(y)) \leq \hat{R}(\theta, \hat{\theta}_{ST}(y)), \\
\hat{\theta}_{ST} & \text{otherwise} 
\end{cases} \]

How does the hybrid estimator perform?
Loss functions

\[ y = \theta + w \]

\[ L(\theta, \hat{\theta}_{EB}(y)) = \frac{1}{n} \| \theta - \hat{\theta}_{EB}(y) \|^2, \]

\[ L(\theta, \hat{\theta}_{ST}(y)) = \frac{1}{n} \| \theta - \hat{\theta}_{ST}(y) \|^2, \]

\[ L_{min}(\theta, y) := \min \left\{ L(\theta, \hat{\theta}_{EB}(y)), L(\theta, \hat{\theta}_{ST}(y)) \right\} \]
Theorem

Consider a sequence of $\theta = \theta(n)$ such that $\frac{1}{n} \sum_i \theta_i^4$ is bounded.

1. For any $t > 0$,

$$\mathbb{P} \left( L(\theta, \hat{\theta}_H(y)) \geq L_{\min}(\theta, y) + t + \kappa_n \right) \leq Ke^{-nk \min(t, t^2)}$$

where $\kappa_n = O\left(\frac{1}{\sqrt{n}}\right)$ and $k, K$ are absolute positive constants.
**Theorem**

Consider a sequence of $\theta = \theta(n)$ such that $\frac{1}{n} \sum_i \theta_i^4$ is bounded.

1. For any $t > 0$,

$$\mathbb{P} \left( L(\theta, \hat{\theta}_H(\mathbf{y})) \geq L_{\text{min}}(\theta, \mathbf{y}) + t + \kappa_n \right) \leq Ke^{-nk \min(t, t^2)}$$

where $\kappa_n = O\left(\frac{1}{\sqrt{n}}\right)$ and $k, K$ are absolute positive constants.

2. The risk of the hybrid estimator satisfies

$$R(\theta, \hat{\theta}_H) \leq \min \left\{ R(\theta, \hat{\theta}_{EB}), R(\theta, \hat{\theta}_{ST}) \right\} + O \left(\frac{1}{\sqrt{n}}\right)$$
Concentration inequalities for $\hat{R}(\theta, \hat{\theta}_{EB}(y))$, $L(\theta, \hat{\theta}_{EB}(y))$

$$
P\left( \left| \hat{R}(\theta, \hat{\theta}_{EB}(y)) - R_1(\theta, \hat{\theta}_{EB}) \right| \geq t \right) \leq Ke^{-nk \min(t, t^2)}$$

where $R_1(\theta, \hat{\theta}_{EB})$ is a deterministic quantity such that

$$
\left| R_1(\theta, \hat{\theta}_{EB}) - R(\theta, \hat{\theta}_{EB}) \right| = O\left( \frac{1}{\sqrt{n}} \right)
$$
Concentration inequalities for $\hat{R}(\theta, \hat{\theta}_{EB}(y))$, $L(\theta, \hat{\theta}_{EB}(y))$

$$
P \left( \left| \hat{R}(\theta, \hat{\theta}_{EB}(y)) - R(\theta, \hat{\theta}_{EB}) \right| \geq t \right) \leq Ke^{-nk \min(t,t^2)}$$

where $R_1(\theta, \hat{\theta}_{EB})$ is a deterministic quantity such that

$$
\left| R_1(\theta, \hat{\theta}_{EB}) - R(\theta, \hat{\theta}_{EB}) \right| = O \left( \frac{1}{\sqrt{n}} \right)
$$

$$
P \left( \left| L(\theta, \hat{\theta}_{EB}(y)) - R_2(\theta, \hat{\theta}_{EB}) \right| \geq t \right) \leq Ke^{-nk \min(t,t^2)}$$

where $R_2(\theta, \hat{\theta}_{EB})$ is a deterministic quantity such that

$$
\left| R_2(\theta, \hat{\theta}_{EB}) - R(\theta, \hat{\theta}_{EB}) \right| = O \left( \frac{1}{\sqrt{n}} \right)$$
Application in Compressed Sensing

\[ y = A\theta + w \]

- \( A \) has i.i.d. sub-Gaussian entries with variance \( 1/m \).
- \( w \sim \mathcal{N}(0, \sigma^2 I) \), \( \sigma \) not necessarily known.
- \( \theta \) is \( \eta \)-sparse, \( \eta \) possibly unknown.
Approximate Message Passing (AMP)

\[ y = A\theta + w \]

Starting from \( \theta_0 = 0 \), generate estimates \( \theta_1, \theta_2, \ldots \):

\[ z_t = y - A\theta_t + b_t z_{t-1} \]

\[ \theta_{t+1} = f_t \left( \theta_t + A^T z_t \right) \]

Key AMP facts:

- As \( n \) grows \( \theta_t + A^T z_t \Rightarrow \theta + \tau_t g \), where \( g \sim_{iid} N(0, 1) \)
- \( \hat{\tau}_t^2 := \|z_t^2\|/m \) is a good estimate of \( \tau_t^2 \)
- \( f_t \) estimates sparse \( \theta \) from observation in Gaussian noise

---

[Donoho, Maleki, Montanari ’09], [Rangan ’10], [Krzakala et al. ’12]...
Simulation Results

Plot of $\frac{1}{n}\|\theta_t - \theta\|^2$ vs $t$, $n = 10,000$

$\eta = 0.13$, $\frac{m}{n} = 0.65$, $\sigma = 1$

Non-zero entries drawn $\sim$ Uniform$[-5, 5]$
Simulation Results

Plot of $\frac{1}{n}\|\theta_t - \hat{\theta}\|^2$ vs $t$, $n = 10,000$

$\eta = 0.05$, $\frac{m}{n} = 0.5$, $\sigma = 0.05$

Non-zero entries drawn from $\{1, -1\}$ uniformly at random
Summary

• SURE to pick better of empirical Bayes and soft-thresholding

• Performance guarantee via concentration results on loss function, bounds on risk

• Idea can be extended to pick one of multiple estimators, e.g., empirical Bayes defined via other priors

http://arxiv.org/abs/1707.09161