Shrinkage Estimation in High Dimensions

Pavan Srinath and Ramji Venkataramanan
University of Cambridge

ITA 2016
The Estimation Problem

\( \theta \in \mathbb{R}^n \) is a vector of parameters, to be estimated from an observation \( y \):

\[
y = \theta + w
\]
i.e.,

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

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

The normalized risk of the estimator is

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

The expectation is calculated with the density

\[
p_\theta(y) = (2\pi)^{-n/2} e^{-\frac{\|y-\theta\|^2}{2}}
\]
Maximum Likelihood Estimator

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

The “obvious” estimator \( \hat{\theta}(y) = y \) is also the ML estimator.

\[ R(\theta, \hat{\theta}_{ML}) = 1 \quad \forall \theta \in \mathbb{R}^n \]

But ...
Maximum Likelihood Estimator

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

The “obvious” estimator \(\hat{\theta}(\mathbf{y}) = \mathbf{y}\) is also the ML estimator.

$$R(\theta, \hat{\theta}_{ML}) = 1 \quad \forall \theta \in \mathbb{R}^n$$

But . . .

For \(n > 2\), there are estimators that do \textit{strictly better for all} \(\theta\)

(James-Stein ’61)
James-Stein Estimator

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

\( \hat{\theta}_{JS} \) shrinks each \( y_i \) towards 0. Its risk is

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

\( n = 10, \ \theta_i = c \) for \( i = 1, \ldots, 5 \) and \( \theta_i = -c \) for \( i = 6, \ldots, 10 \)
Intuition

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

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

Why should the estimate of \( \theta_1 \) depend on all the \( y_i \)'s?
Intuition

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

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

Why should the estimate of \( \theta_1 \) depend on all the \( y_i \)'s?

Note that the loss function is \( \frac{1}{n} \left( (\theta_1 - \hat{\theta}_1)^2 + \ldots + (\theta_n - \hat{\theta}_n)^2 \right) \)
Intuition

\[
\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} + \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_n \end{bmatrix}, \quad \omega_i \text{ i.i.d. } \sim \mathcal{N}(0, 1)
\]

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

**Empirical Bayes:** If we assume \( \theta_i \text{ i.i.d. } \sim \mathcal{N}(0, A) \), then

\[
\hat{\theta}_{MMSE} = \left( 1 - \frac{1}{1+A} \right) \mathbf{y}
\]

If we don’t know \( \frac{1}{1+A} \), estimate it by \( \frac{n-2}{\|\mathbf{y}\|^2} \) since \( \mathbb{E} \left[ \frac{n-2}{\|\mathbf{y}\|^2} \right] = \frac{1}{1+A} \)

This gives \( \hat{\theta}_{JS} \), but the surprise is that it beats ML for any \( \theta \)!
The Attracting Vector

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

\( \hat{\theta}_{JS} \) shrinks \( y \) towards the all-zeros vector \( 0 \)
Risk smaller when \( \theta \) closer to \( 0 \), i.e., \( \|\theta\| \) is small
The Attracting Vector

In general, can shrink towards *any* vector, e.g., $\alpha \mathbf{1}$

$$\hat{\theta} = \alpha \mathbf{1} + \left[ 1 - \frac{(n-2)}{\|\mathbf{y} - \alpha \mathbf{1}\|^2} \right] (\mathbf{y} - \alpha \mathbf{1})$$

Risk of $\hat{\theta}$ decreases as $\|\theta - \alpha \mathbf{1}\|$ gets smaller
The Attracting Vector

In general, can shrink towards any vector, e.g., $\alpha \mathbf{1}$

$$
\hat{\theta} = \alpha \mathbf{1} + \left[1 - \frac{(n-2)}{\|\mathbf{y} - \alpha \mathbf{1}\|^2}\right] (\mathbf{y} - \alpha \mathbf{1})
$$

Risk of $\hat{\theta}$ decreases as $\|\theta - \alpha \mathbf{1}\|$ gets smaller

*Lindley’s estimator:*

Based on assumption that $\theta_i$’s lie close to their average $\bar{\theta} (\approx \bar{y})$

$$
\hat{\theta}_L = \bar{y} \mathbf{1} + \left[1 - \frac{(n-3)}{\|\mathbf{y} - \bar{y} \mathbf{1}\|^2}\right] (\mathbf{y} - \bar{y} \mathbf{1}), \quad \bar{y} = \sum_i \frac{y_i}{n}
$$

$\hat{\theta}_L$ has been applied to baseball data, disease incidence data... [Efron-Morris ’75]

Risk reduction of a JS-like estimator over ML is greatest when $\theta$ is close to the attracting vector
Positive-Part Version

\[ \hat{\theta}_L = \bar{y}1 + \left[ 1 - \frac{(n-3)}{\|y - \bar{y}1\|^2} \right]_+ (y - \bar{y}1), \quad \bar{y} = \sum \frac{y_i}{n} \]

When shrinkage factor becomes -ve, replace it by 0
Positive-Part Version

\[ \hat{\theta}_L = \bar{y}1 + \left[ 1 - \frac{(n-3)}{\|y - \bar{y}1\|^2} \right] (y - \bar{y}1), \quad \bar{y} = \sum_i \frac{y_i}{n} \]

When shrinkage factor becomes -ve, replace it by 0

\[ n = 10, \; \theta_i = c \; \text{for all} \; i \]

\[ \hat{\theta}_L \] does best when \( \theta_i \)'s are close to their empirical mean \( \bar{\theta} \)
Another example

\[ n = 10, \theta_i = c \text{ for } 1 \leq i \leq 5, \text{ and } \theta_i = -c \text{ for } 6 \leq i \leq 10 \]

What would be a good attracting vector for this example?
Shrink +ve \( y_i \)'s that are \( > \bar{y} \) towards \( c \), the rest towards \( -c \).

But we don’t know anything about \( \theta \)!
Components of $\theta$ in 2 clusters

Components of $\theta$ in 1 cluster

- In the absence of prior information about $\theta$, how do we choose an attracting vector that is close to $\theta$?
- Can we use the data to pick a good attracting vector tailored to the underlying $\theta$?
Components of $\theta$ in 2 clusters

Components of $\theta$ in 1 cluster

- In the absence of prior information about $\theta$, how do we choose an attracting vector that is close to $\theta$?
- Can we use the data to pick a good attracting vector tailored to the underlying $\theta$?

Idea:

1) Design a good estimator $\hat{\theta}_2$ for $\theta$’s whose components are roughly separable into two clusters
2) Then from $y$, try to infer which is better — $\hat{\theta}_2$ or $\hat{\theta}_L$
A Two-Cluster Estimator

Define two clusters

\[ C_1 := \{y_i \mid y_i > \bar{y}\}, \quad C_2 := \{y_i \mid y_i \leq \bar{y}\}. \]

Shrink \( y_i \)'s in \( C_1 \) towards \( a_1 \), the rest towards \( a_2 \).
A Two-Cluster Estimator

Define two clusters

\[ C_1 := \{ y_i \mid y_i > \bar{y} \}, \quad C_2 := \{ y_i \mid y_i \leq \bar{y} \}. \]

Shrink \( y_i \)'s in \( C_1 \) towards \( a_1 \), the rest towards \( a_2 \).

Attracting vector:

\[ \nu_2 = a_1 \begin{bmatrix} 1_{\{y_1 > \bar{y}\}} \\ \vdots \\ 1_{\{y_n > \bar{y}\}} \end{bmatrix} + a_2 \begin{bmatrix} 1_{\{y_1 \leq \bar{y}\}} \\ \vdots \\ 1_{\{y_n \leq \bar{y}\}} \end{bmatrix} \]

The estimator is

\[ \hat{\theta}_2 = \nu_2 + \left[ 1 - \frac{n}{\| \mathbf{y} - \nu_2 \|^2} \right]_+ (\mathbf{y} - \nu_2) \]

How to choose \( a_1 \) and \( a_2 \)?
The Ideal Attractors

Attracting vector $\nu_2$ lies in a 2-d subspace spanned by

\[
\begin{bmatrix}
1_{\{y_1 > \bar{y}\}} \\
\vdots \\
1_{\{y_n > \bar{y}\}}
\end{bmatrix}
\quad \text{and} \quad
\begin{bmatrix}
1_{\{y_1 \leq \bar{y}\}} \\
\vdots \\
1_{\{y_n \leq \bar{y}\}}
\end{bmatrix}
\]

Key Idea: Best attractor is the vector in this subspace that is closest to $\theta \Rightarrow$ Projection of $\theta$ onto this subspace:
The Ideal Attractors

Attracting vector $\nu_2$ lies in a 2-d subspace spanned by

$$
\begin{bmatrix}
1\{y_1 > \bar{y}\} \\
\vdots \\
1\{y_n > \bar{y}\}
\end{bmatrix}
$$

and

$$
\begin{bmatrix}
1\{y_1 \leq \bar{y}\} \\
\vdots \\
1\{y_n \leq \bar{y}\}
\end{bmatrix}
$$

**Key Idea**: Best attractor is the vector in this subspace that is closest to $\theta$ $\Rightarrow$ Projection of $\theta$ onto this subspace:

$$
\nu_2^* = a_1^* \begin{bmatrix}
1\{y_1 > \bar{y}\} \\
\vdots \\
1\{y_n > \bar{y}\}
\end{bmatrix} + a_2^* \begin{bmatrix}
1\{y_1 \leq \bar{y}\} \\
\vdots \\
1\{y_n \leq \bar{y}\}
\end{bmatrix}
$$

$$
a_1^* = \frac{\sum_{i=1}^{n} \theta_i 1\{y_i > \bar{y}\}}{\sum_{i=1}^{n} 1\{y_i > \bar{y}\}}, \quad a_2^* = \frac{\sum_{i=1}^{n} \theta_i 1\{y_i \leq \bar{y}\}}{\sum_{i=1}^{n} 1\{y_i \leq \bar{y}\}}.
$$

$a_1, a_2$ cannot be computed, but can we estimate them from $y$?
Estimating the Attractors

\[
a_1 = \frac{\sum_{i=1}^{n} y_i 1\{y_i > \bar{y}\} - \frac{1}{2\delta} \sum_{i=0}^{n} 1\{|y_i - \bar{y}| \leq \delta\}}{\sum_{i=1}^{n} 1\{y_i > \bar{y}\}}
\]

\[
a_2 = \frac{\sum_{i=1}^{n} y_i 1\{y_i \leq \bar{y}\} + \frac{1}{2\delta} \sum_{i=0}^{n} 1\{|y_i - \bar{y}| \leq \delta\}}{\sum_{i=1}^{n} 1\{y_i \leq \bar{y}\}}
\]

\(\delta\) is a constant that should be chosen small but \(\gg \frac{1}{\sqrt{n}}\)

**Concentration of \(a_1, a_2\)**

For any \(0 < \epsilon < 1\)

\[
P (|a_1 - a_1^* - \kappa_1 \delta + o(\delta)| \geq \epsilon) \leq Ke^{-nk\epsilon^2}
\]

\[
P (|a_2 - a_2^* - \kappa_2 \delta + o(\delta)| \geq \epsilon) \leq K' e^{-nk'\epsilon^2}
\]
Risk of Two-Cluster Estimator

\[ \hat{\theta}_2 = \nu_2 + \left[ 1 - \frac{n}{\|y - \nu_2\|^2} \right] (y - \nu_2) \]

**Theorem**: The loss function of the two-cluster estimator \( \hat{\theta}_2 \) satisfies:

1. For any \( 0 < \epsilon < 1 \),
   \[
   \mathbb{P} \left( \left| \frac{1}{n} \|\theta - \hat{\theta}_2\|^2 - \left[ \min \left( \beta_n, \frac{\beta_n}{\alpha_n + 1} \right) + \kappa_n \delta + o(\delta) \right] \right| \geq \epsilon \right) \leq K \epsilon^{-n \kappa \epsilon^2}
   \]

   where \( \alpha_n, \beta_n \) are explicit constants that depend on \( \theta \).
Risk of Two-Cluster Estimator

\[ \hat{\theta}_2 = \nu_2 + \left[1 - \frac{n}{\|y - \nu_2\|^2}\right] (y - \nu_2) \]

**Theorem:** The loss function of the two-cluster estimator \( \hat{\theta}_2 \) satisfies:

(1) For any \( 0 < \epsilon < 1 \),

\[ \mathbb{P}\left( \left| \frac{1}{n} \|\theta - \hat{\theta}_2\|^2 - \left[ \min \left( \beta_n, \frac{\beta_n}{\alpha_n + 1} \right) + \kappa_n \delta + o(\delta) \right] \right| \geq \epsilon \right) \leq K \epsilon^{-nk\epsilon^2} \]

where \( \alpha_n, \beta_n \) are explicit constants that depend on \( \theta \)

(2) For a sequence of \( \theta \) with increasing dimension \( n \), if \( \limsup_{n \to \infty} \frac{\|\theta\|^2}{n} < \infty \), we have

\[ \lim_{n \to \infty} \left| \frac{1}{n} R(\theta, \hat{\theta}_2) - \left[ \min \left( \beta_n, \frac{\beta_n}{\alpha_n + 1} \right) + \kappa_n \delta + o(\delta) \right] \right| = 0 \]
Risk of Lindley’s Estimator

Note that $\hat{\theta}_L$ is a one-cluster estimator: Best attractor in 1d subspace spanned by $\mathbf{1}$ is $\bar{\theta}\mathbf{1}$; $\hat{\theta}_L$ approximates it by $\bar{y}\mathbf{1}

**Corollary**: The loss function of Lindley’s estimator satisfies:

1. For any $0 < \epsilon < 1$,

$$P \left( \left| \left\| \theta - \hat{\theta}_L \right\|^2 - \frac{\rho_n}{\rho_n + 1} \right| \geq \epsilon \right) \leq K e^{-nk\epsilon^2},$$

where

$$\rho_n = \frac{\left\| \theta - \bar{\theta}\mathbf{1} \right\|^2}{n}.$$

2. If $\limsup_{n \to \infty} \frac{\|\theta\|^2}{n} < \infty$, we have

$$\lim_{n \to \infty} \left| \frac{1}{n} R \left( \theta, \hat{\theta}_{JS_1} \right) - \frac{\rho_n}{\rho_n + 1} \right| = 0.$$
Picking the Better Estimator

Depending on $\theta$, either $\hat{\theta}_2$ or $\hat{\theta}_L$ may have lower risk.

*Expect $\hat{\theta}_2$ to be better*

- For $\theta$ whose components are roughly separable into two clusters, expect $\hat{\theta}_2$ to have lower risk than $\hat{\theta}_L$.
- For $\theta$ whose components are clustered around one value, expect $\hat{\theta}_L$ to have lower risk.

How to pick the better estimator?
Picking the Better Estimator

Depending on $\theta$, either $\hat{\theta}_2$ or $\hat{\theta}_L$ may have lower risk

---

Expect $\hat{\theta}_2$ to be better

- For $\theta$ whose components are roughly separable into two clusters, expect $\hat{\theta}_2$ to have lower risk than $\hat{\theta}_L$
- For $\theta$ whose components are clustered around one value, expect $\hat{\theta}_L$ to have lower risk

---

How to pick the better estimator?

IDEA: Estimate the loss of each estimator from the data
Hybrid Estimator

Loss Estimates:

\[
\hat{L}(\theta, \hat{\theta}_L) = \left(1 - \frac{1}{g\left(\|y - \bar{y}\|_2^2/n\right)}\right)
\]

\[
\hat{L}(\theta, \hat{\theta}_2) = \frac{1}{n} \|y - \nu_2\|_2^2 - 1 + \frac{1}{n\delta} \left(\sum_{i=0}^{n} 1_{\{|y_i - \bar{y}| \leq \delta\}}\right) \frac{(a_1 - a_2)}{g(\|y - \nu_2\|_2^2/n)}
\]

where \( g(x) = \max\{x, 1\} \)

Choose

\[
\hat{\theta}_{hyb} = \begin{cases} 
\hat{\theta}_L & \text{if } \hat{L}(\theta, \hat{\theta}_L) \leq \hat{L}(\theta, \hat{\theta}_2), \\
\hat{\theta}_2 & \text{otherwise}
\end{cases}
\]

Different from approach in [George '86]: convex combinations of multiple shrinkage estimators with fixed attracting subspaces.
Performance of Hybrid Estimator

**Theorem:** The loss function of the hybrid JS-estimator satisfies:

(1) For any $0 < \epsilon < 1$,

$$
\mathbb{P} \left( \left| \frac{1}{n} \| \theta - \hat{\theta}_{hyb} \|_2^2 - \min \left\{ \frac{1}{n} \| \theta - \hat{\theta}_L \|_2^2, \frac{1}{n} \| \theta - \hat{\theta}_2 \|_2^2 \right\} \right| \geq \epsilon \right) \\
\leq Ke^{-nk\epsilon^2}
$$

where $K$ and $k$ are positive constants.

(2) For a sequence of $\theta$ with increasing dimension $n$, if

$$
\limsup_{n \to \infty} \frac{\| \theta \|_2^2}{n} < \infty,
$$

we have

$$
\lim_{n \to \infty} \left| \frac{1}{n} R(\theta, \hat{\theta}_{hyb}) - \min \left\{ \frac{1}{n} R(\theta, \hat{\theta}_L), \frac{1}{n} R(\theta, \hat{\theta}_2) \right\} \right| = 0
$$

Proof involves showing $\hat{L}(\theta, \hat{\theta}_L), \hat{L}(\theta, \hat{\theta}_2)$ concentrate around

$\frac{1}{n} \| \theta - \hat{\theta}_L \|_2^2, \frac{1}{n} \| \theta - \hat{\theta}_2 \|_2^2$, respectively.
θ_i’s arranged in 2 clusters, one centered at τ, the other at −τ
Each cluster has width \( \frac{\tau}{2} \), placement of points within a cluster is uniformly random
Simulation Results

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

\( n = 1000 \)

- Regular JS-estimator
- Lindley’s estimator
- Two-cluster JS-estimator
- Hybrid JS-estimator
- ML-estimator

- \( \theta_i \)'s arranged in 2 clusters, one centered at \( \tau \), the other at \( -\tau \)
- Each cluster has width \( \frac{\tau}{2} \), placement of points within a cluster is uniformly random
Simulation Results

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

\( n = 1000 \)

\( \theta_i \)'s are uniformly placed between \(-\tau\) and \(\tau\)
Summary

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

To achieve significant risk reduction over ML:

- Shrinkage estimator needs attracting vector \( \nu \) that’s close to \( \theta \)
- Can we find a good \( \nu(y) \), without any knowledge about \( \theta \)?

- Proposed estimator infers clustering structure of \( \theta \) and picks a good target subspace for the attractor
- Can test for up to \( L \)-clusters for any integer \( L \geq 2 \), and pick the best one based on loss estimate.
- Provided concentration results for loss function, convergence results for risk

Future Work: Test performance on real data sets,
More general multi-dimn. target subspaces than cluster-based ones