Cluster-Seeking Shrinkage Estimators

K. Pavan Srinath
University of Cambridge, UK
Email: pk423@cam.ac.uk

Ramji Venkataramanan
University of Cambridge, UK
Email: ramji.v@eng.cam.ac.uk

***Abstract*—This paper considers the problem of estimating a high-dimensional vector $\boldsymbol{\theta} \in \mathbb{R}^n$ from a noisy one-time observation. The noise vector is assumed to be i.i.d. Gaussian with known variance. For the squared-error loss function, the James-Stein (JS) estimator is known to dominate the simple maximum-likelihood (ML) estimator when the dimension n exceeds two. The JS-estimator shrinks the observed vector towards the origin, and the risk reduction over the ML-estimator is greatest for $\boldsymbol{\theta}$ that lie close to the origin. JS-estimators can be generalized to shrink the data towards any target subspace. Such estimators also dominate the ML-estimator, but the risk reduction is significant only when $\boldsymbol{\theta}$ lies close to the subspace. This leads to the question: in the absence of prior information about $\boldsymbol{\theta}$, how do we design estimators that give significant risk reduction over the ML-estimator for a wide range of $\boldsymbol{\theta}$?**

In this paper, we attempt to infer the structure of $\boldsymbol{\theta}$ from the observed data in order to construct a good attracting subspace for the shrinkage estimator. We provide concentration results for the squared-error loss and convergence results for the risk of the proposed estimators, as well as simulation results to support the claims. The estimators give significant risk reduction over the ML-estimator for a wide range of $\boldsymbol{\theta}$, particularly for large n.

I. Introduction

Consider the problem of estimating a vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^n$ from a noisy observation $\mathbf{y}$ of the form

$$\mathbf{y} = \boldsymbol{\theta} + \mathbf{w},$$

where the noise vector $\mathbf{w} \in \mathbb{R}^n$ is distributed as $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, i.e., its components are i.i.d. Gaussian random variables with mean zero and variance σ^2. We emphasize that $\boldsymbol{\theta}$ is deterministic, so the joint probability density function of $\mathbf{y} = [y_1, \dots, y_n]^T$ for a given $\boldsymbol{\theta}$ is

$$p_{\boldsymbol{\theta}}(\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{\|\mathbf{y}-\boldsymbol{\theta}\|^2}{2\sigma^2}}. \qquad (1)$$

The performance of an estimator $\hat{\boldsymbol{\theta}}$ is measured using the squared-error loss function given by $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{y})) := \|\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}\|^2$, where $\|\cdot\|$ denotes the Euclidean norm. The *risk* of the estimator for a given $\boldsymbol{\theta}$ is the expected value of the loss function:

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) := \mathbb{E}\left[\|\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}\|^2\right],$$

where the expectation is computed using the density in (1). The *normalized risk* is $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})/n$.

Applying the maximum-likelihood (ML) criterion to (1) yields the ML-estimator $\hat{\boldsymbol{\theta}}_{ML} = \mathbf{y}$. The ML-estimator is an unbiased estimator, and its risk is $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ML}) = n\sigma^2$. The goal of this paper is to design estimators that give significant risk reduction over $\hat{\boldsymbol{\theta}}_{ML}$ for a wide range of $\boldsymbol{\theta}$, without prior assumptions about its structure.

In 1961 James and Stein published a surprising result [1], proposing an estimator that uniformly achieves lower risk than $\hat{\boldsymbol{\theta}}_{ML}$ for any $\boldsymbol{\theta} \in \mathbb{R}^n$, for $n \geq 3$. Their estimator $\hat{\boldsymbol{\theta}}_{JS}$ is given by

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

and its risk is [2, Chapter 5, Thm. 5.1]

$$R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS}\right) = n\sigma^2 - (n-2)^2\sigma^4\mathbb{E}\left[\frac{1}{\|\mathbf{y}\|^2}\right]. \qquad (3)$$

Hence for $n \geq 3$, $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS}) < R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ML}) = n\sigma^2$, $\forall \boldsymbol{\theta} \in \mathbb{R}^n$. An estimator $\hat{\boldsymbol{\theta}}_1$ is said to *dominate* another estimator $\hat{\boldsymbol{\theta}}_2$ if $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_1) \leq R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_2)$, $\forall \boldsymbol{\theta} \in \mathbb{R}^n$, with the inequality being strict for at least one $\boldsymbol{\theta}$. Thus, the James-Stein estimator (JS-estimator) dominates the ML-estimator. Unlike the ML-estimator, the JS-estimator is non-linear and biased. However, the risk reduction over the ML-estimator can be significant, making it an attractive option in many situations—see, for example, [3]. By evaluating the expression in (3), it can be shown that the risk of the JS-estimator depends on $\boldsymbol{\theta}$ only via $\|\boldsymbol{\theta}\|$ [1], and decreases with $\|\boldsymbol{\theta}\|$.

The JS-estimator belongs to a class of estimators called *shrinkage* estimators. In particular, $\hat{\boldsymbol{\theta}}_{JS}$ shrinks each element of $\mathbf{y}$ towards the origin. Extending this idea, JS-like estimators can be defined by shrinking $\mathbf{y}$ towards any vector, or more generally, towards a target subspace $\mathbb{V} \subset \mathbb{R}^n$. Let $P_{\mathbb{V}}(\mathbf{y})$ denote the projection of $\mathbf{y}$ onto $\mathbb{V}$, so that $\|\mathbf{y} - P_{\mathbb{V}}(\mathbf{y})\|^2 = \min_{\mathbf{v}\in\mathbb{V}} \|\mathbf{y}-\mathbf{v}\|^2$. Then the JS-estimator that shrinks $\mathbf{y}$ towards the subspace $\mathbb{V}$ is

$$\hat{\boldsymbol{\theta}} = P_{\mathbb{V}}(\mathbf{y}) + \left[1 - \frac{(n-d-2)\sigma^2}{\|\mathbf{y} - P_{\mathbb{V}}(\mathbf{y})\|^2}\right](\mathbf{y} - P_{\mathbb{V}}(\mathbf{y})), \qquad (4)$$

where d is the dimension of $\mathbb{V}$.[1] A classic example of such an estimator is Lindley's estimator [4], which shrinks $\mathbf{y}$ towards the one-dimensional subspace defined by the all-ones vector $\mathbf{1}$. It is given by

$$\hat{\boldsymbol{\theta}}_L = \bar{y}\mathbf{1} + \left[1 - \frac{(n-3)\sigma^2}{\|\mathbf{y} - \bar{y}\mathbf{1}\|^2}\right](\mathbf{y} - \bar{y}\mathbf{1}), \qquad (5)$$

where $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i$ is the empirical mean of $\mathbf{y}$.

It can be shown that the different variants of the JS-estimator such as (2), (4) and (5) all dominate the ML-estimator. Further, all JS-estimators share the following key property [5]–[7]: **the smaller the Euclidean distance between $\boldsymbol{\theta}$ and the attracting vector, the smaller the risk**.

[1] The dimension n has to be greater than $d + 2$ for the estimator to achieve lower risk than $\hat{\boldsymbol{\theta}}_{ML}$.

978-1-5090-1806-2/16/$31.00 ©2016 IEEE

For $\hat{\boldsymbol{\theta}}_{JS}$ in (2), the attracting vector is $\mathbf{0}$, and the risk reduction over $\hat{\boldsymbol{\theta}}_{ML}$ is greatest when $\boldsymbol{\theta}$ is close to the origin. Similarly, if the components of $\boldsymbol{\theta}$ are clustered around some value c, a JS-estimator with attracting vector $c\mathbf{1}$ would give significant risk reduction over $\hat{\boldsymbol{\theta}}_{ML}$. One motivation for Lindley's estimator in (7) comes from a guess that the components of $\boldsymbol{\theta}$ are close to its empirical mean $\bar{\theta}$ — since we do not know $\bar{\theta}$, we approximate it by $\bar{y}$ and use the attracting vector $\bar{y}\mathbf{1}$.

The risk reduction obtained by using a JS-like shrinkage estimator over $\hat{\boldsymbol{\theta}}_{ML}$ crucially depends on the choice of attracting vector. To achieve significant risk reduction for a wide range of $\boldsymbol{\theta}$, in this paper, we aim to infer the structure of $\boldsymbol{\theta}$ from the data $\mathbf{y}$ and choose attracting vectors tailored to this structure. The main idea is to partition $\mathbf{y}$ into clusters, and shrink the components in each cluster towards a common element (attractor). Both the number of clusters and the attractor for each cluster are to be determined based on $\mathbf{y}$.

As a motivating example, consider a $\boldsymbol{\theta}$ in which half the components are equal to $\|\boldsymbol{\theta}\|/\sqrt{n}$ and the other half are equal to $-\|\boldsymbol{\theta}\|/\sqrt{n}$. An ideal JS-estimator would shrink the y_i's corresponding to $\theta_i = \|\boldsymbol{\theta}\|/\sqrt{n}$ towards the attractor $\|\boldsymbol{\theta}\|/\sqrt{n}$, and the remaining observations towards $-\|\boldsymbol{\theta}\|/\sqrt{n}$. Such an estimator would give handsome gains over $\hat{\boldsymbol{\theta}}_{ML}$ for all $\boldsymbol{\theta}$ with the above structure. On the other hand, if $\boldsymbol{\theta}$ is such that all its components are equal (to $\bar{\theta}$), Lindley's estimator $\hat{\boldsymbol{\theta}}_L$ is an excellent choice, with significantly smaller risk than $\hat{\boldsymbol{\theta}}_{ML}$ for all values of $\|\boldsymbol{\theta}\|$.

We would like an intelligent estimator that can correctly distinguish between different $\boldsymbol{\theta}$ structures (such as the two above) and choose an appropriate attracting vector, based only on $\mathbf{y}$. We propose such estimators in Sections II and III of this paper. For reasonably large n, these estimators choose a good attracting subspace tailored to the structure of $\boldsymbol{\theta}$, and use an approximation of the best attracting vector within the subspace. The main contributions of our paper are as follows.

- In Section II, we construct a two-cluster JS-estimator, and provide concentration results for the squared-error loss, and asymptotic convergence results for its risk. This estimator is shown to provide significant risk reduction over Lindley's estimator and the regular JS-estimator when the components of $\boldsymbol{\theta}$ can be approximately separated into two clusters.
- In Section III , we present a hybrid JS-estimator that, for any $\boldsymbol{\theta}$ and for large n, has risk close to the minimum of that of Lindley's estimator and the proposed two-cluster JS-estimator.
- In Section IV, we provide simulation results that support the theoretical results on the loss function. The simulations indicate that the hybrid JS-estimator can give significant risk reduction even for moderately large n, e.g. $n = 50$.

The ideas in this paper are generalized in [8] to construct hybrid JS-estimators with multiple clusters. Due to space constraints, the proofs of the theorems as well as extensive simulation results are given in [8].

A. Positive-Part Shrinkage Estimators

Noting that the shrinkage factor multiplying $\mathbf{y}$ in (2) could be negative, Stein proposed the following positive-part JS-estimator [1]:

$$\hat{\boldsymbol{\theta}}_{JS_+} = \left[1 - \frac{(n-2)\sigma^2}{\|\mathbf{y}\|^2}\right]_+ \mathbf{y}, \tag{6}$$

where X_+ denotes $\max(0, X)$. Similarly, the positive-part Lindley's estimator is given by

$$\hat{\boldsymbol{\theta}}_{L_+} = \bar{y}\mathbf{1} + \left[1 - \frac{(n-3)\sigma^2}{\|\mathbf{y} - \bar{y}\mathbf{1}\|^2}\right]_+ (\mathbf{y} - \bar{y}\mathbf{1}). \tag{7}$$

Baranchik [9] proved that $\hat{\boldsymbol{\theta}}_{JS_+}$ dominates $\hat{\boldsymbol{\theta}}_{JS}$, and his result also proves that $\hat{\boldsymbol{\theta}}_{L_+}$ dominates $\hat{\boldsymbol{\theta}}_L$. However, for large n, the shrinkage factor is positive with high probability, hence the positive-part estimators are nearly always identical to their corresponding non positive-part versions. Estimators that dominate $\hat{\boldsymbol{\theta}}_{JS_+}$ are discussed in [10], [11]. Henceforth in this paper, by regular JS-estimator and Lindley's estimator, we mean $\hat{\boldsymbol{\theta}}_{JS_+}$ and $\hat{\boldsymbol{\theta}}_{L_+}$, respectively. The estimators proposed in this paper are also positive part estimators.

Notation: Bold lowercase letters are used to denote vectors, and plain lowercase letters for their entries. For example, the entries of $\mathbf{y} \in \mathbb{R}^n$ are y_i, $i = 1, \cdots, n$. All vectors have length n and are column vectors, unless otherwise mentioned. We use $1_{\{\mathcal{E}\}}$ to denote the indicator function of an event $\mathcal{E}$. The Q-function is given by $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})dx$, and $Q^c(x)$ denotes $1 - Q(x)$. For a random variable X, X_+ denotes $\max(0, X)$. For real functions $f(x)$ and $g(x)$, the notation $f(x) = o(g(x))$ means that $\lim_{x \to 0}[f(x)/g(x)] = 0$.

Finally, we use the following shorthand for concentration inequalities. Let $\{X_n(\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^n\}_{n=1}^\infty$ be a sequence of random variables. The notation $X_n(\boldsymbol{\theta}) \doteq X$, where X is either a random variable or a constant, means that for any $\epsilon > 0$,

$$\mathbb{P}(|X_n(\boldsymbol{\theta}) - X| \geq \epsilon) \leq K e^{-\frac{nk\min(\epsilon^2, 1)}{\max(\|\boldsymbol{\theta}\|^2/n, 1)}}, \tag{8}$$

where K and k are positive constants that do not depend on n or $\boldsymbol{\theta}$. The exact values of K and k are not specified.

II. A two-cluster James-Stein estimator

Recall the example in Section I where $\boldsymbol{\theta}$ has half its components equal to $\|\boldsymbol{\theta}\|/\sqrt{n}$, and the other half equal to $\|\boldsymbol{\theta}\|/\sqrt{n}$. Ideally, we would to shrink the y_i's corresponding to the first group towards $\|\boldsymbol{\theta}\|/\sqrt{n}$, and the remaining points towards $-\|\boldsymbol{\theta}\|/\sqrt{n}$. However, without an oracle, we cannot accurately guess which attractor each y_i should be shrunk towards. We would like to obtain an estimator that identifies separable clusters in $\mathbf{y}$, constructs a suitable attractor for each cluster, and shrinks the y_i in each cluster towards its attractor.

We start by dividing the observed data into two clusters based on a separating point $s_{\mathbf{y}}$, which is obtained from $\mathbf{y}$. A natural choice for the $s_{\mathbf{y}}$ would be the empirical mean $\bar{\theta}$; since this is unknown we use $s_{\mathbf{y}} = \bar{y}$. Define the clusters

$$\mathcal{C}_1 := \{y_i,\ 1 \leq i \leq n \mid y_i > \bar{y}\},$$
$$\mathcal{C}_2 := \{y_i,\ 1 \leq i \leq n \mid y_i \leq \bar{y}\}.$$

The points in $\mathcal{C}_1$ and $\mathcal{C}_2$ will be shrunk towards attractors $a_1 := f_1(\mathbf{y})$ and $a_2 := f_2(\mathbf{y})$, respectively, where the functions $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ are defined later in this section (see (12) and (13)). Thus the attracting vector is

$$\boldsymbol{\nu}_2 := a_1 \begin{bmatrix} 1_{\{y_1>\bar{y}\}} \\ 1_{\{y_2>\bar{y}\}} \\ \vdots \\ 1_{\{y_n>\bar{y}\}} \end{bmatrix} + a_2 \begin{bmatrix} 1_{\{y_1\leq\bar{y}\}} \\ 1_{\{y_2\leq\bar{y}\}} \\ \vdots \\ 1_{\{y_n\leq\bar{y}\}} \end{bmatrix}, \tag{9}$$

and the proposed estimator is

$$\hat{\boldsymbol{\theta}}_{JS_2} = \boldsymbol{\nu}_2 + \left[1 - \frac{n\sigma^2}{\|\mathbf{y}-\boldsymbol{\nu}_2\|^2}\right]_+ (\mathbf{y} - \boldsymbol{\nu}_2). \tag{10}$$

The attracting vector $\boldsymbol{\nu}_2$ in (9) lies in a two-dimensional subspace defined by the orthogonal vectors $[1_{\{y_1>\bar{y}\}}, \cdots, 1_{\{y_n>\bar{y}\}}]^T$ and $[1_{\{y_1\leq\bar{y}\}}, \cdots, 1_{\{y_n\leq\bar{y}\}}]^T$. To derive the values of a_1 and a_2 in (9), it is useful to compare $\boldsymbol{\nu}_2$ to the attracting vector of Lindley's estimator in (7). Recall that Lindley's attracting vector lies in the one-dimensional subspace spanned by $\mathbf{1}$. The vector lying in this subspace that is closest in Euclidean distance to $\boldsymbol{\theta}$ is its projection $\bar{\theta}\mathbf{1}$. Since $\bar{\theta}$ is unknown, we use the approximation $\bar{y}$ to define the attracting vector $\bar{y}\mathbf{1}$.

Analogously, the vector in the two-dimensional subspace defined by (9) that is closest to $\boldsymbol{\theta}$ is the projection of $\boldsymbol{\theta}$ onto this subspace. Computing this projection, the desired values for a_1, a_2 are found to be

$$a_1^{des} = \frac{\sum_{i=1}^n \theta_i 1_{\{y_i>\bar{y}\}}}{\sum_{i=1}^n 1_{\{y_i>\bar{y}\}}}, \quad a_2^{des} = \frac{\sum_{i=1}^n \theta_i 1_{\{y_i\leq\bar{y}\}}}{\sum_{i=1}^n 1_{\{y_i\leq\bar{y}\}}}. \tag{11}$$

As the θ_i's are not available, we define the attractors a_1, a_2 as approximations of a_1^{des}, a_2^{des}, obtained using the following concentration results.

Lemma 1. *We have*

$$\frac{1}{n}\sum_{i=1}^n y_i 1_{\{y_i>\bar{y}\}} \doteq \frac{1}{n}\sum_{i=1}^n \theta_i 1_{\{y_i>\bar{y}\}} + \frac{\sigma}{n\sqrt{2\pi}}\sum_{i=1}^n e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}},$$
$$\frac{1}{n}\sum_{i=1}^n y_i 1_{\{y_i\leq\bar{y}\}} \doteq \frac{1}{n}\sum_{i=1}^n \theta_i 1_{\{y_i>\bar{y}\}} - \frac{\sigma}{n\sqrt{2\pi}}\sum_{i=1}^n e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}}.$$

Recall that the symbol '$\doteq$' is shorthand for a concentration inequality of the form (8). The proof of the lemma is given in [8]. Using Lemma 1, we can obtain estimates for a_1^{des}, a_2^{des} in (11) provided we have an estimate for the term $\frac{\sigma}{n\sqrt{2\pi}}\sum_{i=1}^n e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}}$. This is achieved via the following concentration result.

Lemma 2. *Fix $\delta > 0$. Then for any $\epsilon > 0$, we have*

$$\mathbb{P}\left(\left|\frac{\sigma^2}{2n\delta}\sum_{i=0}^n 1_{\{|y_i-\bar{y}|\leq\delta\}} - \frac{\sigma}{n\sqrt{2\pi}}\sum_{i=0}^n e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}} + \kappa_n\delta\right| \geq \epsilon\right) \leq 10e^{-nk\epsilon^2},$$

where k is some positive constant and $|\kappa_n| \leq \frac{1}{\sqrt{2\pi e}}$.

The proof is given in [8]. Henceforth in this paper, κ_n is used to denote a generic bounded constant that is a coefficient of δ in expressions of the form $f(\delta) = a + \kappa_n\delta + o(\delta)$ where a is some constant. The exact value of κ_n is not needed.

Using Lemmas 1 and 2, the two attractors are defined to be

$$a_1 = \frac{\sum_{i=1}^n y_i 1_{\{y_i>\bar{y}\}} - \frac{\sigma^2}{2\delta}\sum_{i=0}^n 1_{\{|y_i-\bar{y}|\leq\delta\}}}{\sum_{i=1}^n 1_{\{y_i>\bar{y}\}}}, \tag{12}$$

$$a_2 = \frac{\sum_{i=1}^n y_i 1_{\{y_i\leq\bar{y}\}} + \frac{\sigma^2}{2\delta}\sum_{i=0}^n 1_{\{|y_i-\bar{y}|\leq\delta\}}}{\sum_{i=1}^n 1_{\{y_i\leq\bar{y}\}}}. \tag{13}$$

With $\delta > 0$ chosen to be a small positive number, this completes the specification of the attracting vector in (9), and hence the two-cluster JS-estimator in (10).

Note 1. *Note that $\boldsymbol{\nu}_2$ is dependent not just on $\mathbf{y}$ but also on δ through the two attractors a_1 and a_2. Further, while δ can be chosen to be arbitrarily small, from a design point of view, the proof of Lemma 2 indicates that δ should be much larger than $1/\sqrt{n}$.*

We now present the first main result of the paper.

Theorem 1. *The loss function of the two-cluster JS-estimator in* (10) *satisfies the following:*

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

$$\mathbb{P}\left(\left|\frac{1}{n}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_2}\|^2 - \left[\min\left(\beta_n, \frac{\beta_n\sigma^2}{\alpha_n+\sigma^2}\right) + \kappa_n\delta + o(\delta)\right]\right| \geq \epsilon\right) \leq Ke^{-\frac{nk\min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}}, \tag{14}$$

where α_n, β_n are given by (16) *and* (17) *below, and K, k are positive constants.*

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

$$\lim_{n\to\infty}\left|\frac{1}{n}R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2}) - \left[\min\left(\beta_n, \frac{\beta_n\sigma^2}{\alpha_n+\sigma^2}\right) + \kappa_n\delta + o(\delta)\right]\right| = 0. \tag{15}$$

The constants α_n, β_n are given by

$$\alpha_n := \frac{\|\boldsymbol{\theta}\|^2}{n} - \frac{c_1^2}{n}\sum_{i=1}^n Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right) - \frac{c_2^2}{n}\sum_{i=1}^n Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right) - \left(\frac{2\sigma}{n\sqrt{2\pi}}\right)\left(\sum_{i=1}^n e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}}\right)(c_1 - c_2), \tag{16}$$

$$\beta_n := \frac{\|\boldsymbol{\theta}\|^2}{n} - \frac{c_1^2}{n}\sum_{i=1}^n Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right) - \frac{c_2^2}{n}\sum_{i=1}^n Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right), \tag{17}$$

where

$$c_1 := \frac{\sum_{i=1}^n \theta_i Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)}{\sum_{i=1}^n Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)}, \quad c_2 := \frac{\sum_{i=1}^n \theta_i Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)}{\sum_{i=1}^n Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)}. \tag{18}$$

The proof of the theorem is given in [8]. The proof also leads to the following result on the performance of Lindley's positive-part estimator $\hat{\boldsymbol{\theta}}_{L_+}$ given by (7).

Corollary 1. *The loss function of the positive-part Lindley's estimator satisfies, for any $\epsilon > 0$,*

$$\mathbb{P}\left(\left| \frac{1}{n} \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{L_+}\|^2 - \frac{\rho_n \sigma^2}{\rho_n + \sigma^2} \right| \geq \epsilon \right) \leq K e^{-nk \min(\epsilon^2, 1)},$$

where K and k are positive constants, and

$$\rho_n := \frac{\|\boldsymbol{\theta} - \bar{\theta}\mathbf{1}\|^2}{n}. \tag{19}$$

III. Hybrid James-Stein Estimator with up to Two Clusters

Depending on the underlying $\boldsymbol{\theta}$, either the positive-part Lindley estimator $\hat{\boldsymbol{\theta}}_{L_+}$ or the two-cluster estimator $\hat{\boldsymbol{\theta}}_{JS_2}$ could have a smaller loss (cf. Theorem 1 and Corollary 1). So we would like an estimator that selects the better among $\hat{\boldsymbol{\theta}}_{L_+}$ and $\hat{\boldsymbol{\theta}}_{JS_2}$ for the $\boldsymbol{\theta}$ in context. Note that this approach is different from the ones in [6], [7] and [12] which consider convex combinations of shrinkage estimators with the weights (non-zero) either prespecified or derived from the data. To this end, we estimate the loss of $\hat{\boldsymbol{\theta}}_{L_+}$ and $\hat{\boldsymbol{\theta}}_{JS_2}$ based on $\mathbf{y}$. Based on these loss estimates, denoted by $\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{L_+})$ and $\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2})$ respectively, we define a hybrid estimator as

$$\hat{\boldsymbol{\theta}}_{JS_H} = \gamma \hat{\boldsymbol{\theta}}_{L_+} + (1 - \gamma) \hat{\boldsymbol{\theta}}_{JS_2}, \tag{20}$$

where $\hat{\boldsymbol{\theta}}_{L_+}$ and $\hat{\boldsymbol{\theta}}_{JS_2}$ are respectively given by (7), (10), and

$$\gamma = \begin{cases} 1 & \text{if} \quad \frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{L_+}) \leq \frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2}), \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

The loss function estimates $\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{L_+})$ and $\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2})$ are obtained as follows. Based on Corollary 1, the loss function of $\hat{\boldsymbol{\theta}}_{L_+}$ can be estimated via an estimate of $\rho_n \sigma^2 / (\rho_n + \sigma^2)$, where ρ_n is given by (19). It is straightforward to check, along the lines of the proof of Theorem 1, that $g(\|\mathbf{y} - \bar{y}\mathbf{1}\|^2/n) \doteq g(\rho_n + \sigma^2) = \rho_n + \sigma^2$, where $g(x) = \max(\sigma^2, x)$. Therefore, an estimate of the normalized loss $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{L_+})/n$ is

$$\frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{L_+}) = \sigma^2 \left(1 - \frac{\sigma^2}{g\left(\|\mathbf{y} - \bar{y}\mathbf{1}\|^2/n \right)} \right). \tag{22}$$

The loss function of the two-cluster estimator $\hat{\boldsymbol{\theta}}_{JS_2}$ can be estimated using Theorem 1, by estimating α_n and β_n defined in (16) and (17), respectively. From Lemma [8], we have

$$\frac{1}{n} \|\mathbf{y} - \boldsymbol{\nu}_2\|^2 \doteq \alpha_n + \sigma^2 + \kappa_n \delta + o(\delta). \tag{23}$$

Further, using the concentration inequalities in Lemmas 1 and 2 in Section II, we can deduce that

$$\begin{aligned} &\frac{1}{n} \|\mathbf{y} - \boldsymbol{\nu}_2\|^2 - \sigma^2 + \frac{\sigma^2}{n\delta} \left(\sum_{i=0}^{n} \mathbf{1}_{\{|y_i - \bar{y}| \leq \delta\}} \right) (a_1 - a_2) \\ &\doteq \beta_n + \kappa_n \delta + o(\delta), \end{aligned} \tag{24}$$

where a_1 and a_2 are defined in (12) and (13), respectively. We now use (23) and (24) to estimate the concentrating value in (14), noting that $\min(\beta_n, (\beta_n\sigma^2)/(\alpha_n + \sigma^2)) = (\beta_n \sigma^2)/g(\alpha_n + \sigma^2)$. This yields

$$\begin{aligned} &\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2})/n \\ &= \frac{\sigma^2 \left(\frac{1}{n} \|\mathbf{y} - \boldsymbol{\nu}_2\|^2 - \sigma^2 + \frac{\sigma^2}{n\delta} \left(\sum_{i=0}^{n} \mathbf{1}_{\{|y_i - \bar{y}| \leq \delta\}} \right) (a_1 - a_2) \right)}{g(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n)}. \end{aligned} \tag{25}$$

The loss function estimates in (22) and (25) complete the specification of the hybrid estimator in (20) and (21). The following theorem characterizes the loss function of the hybrid estimator, by showing that the loss estimates in (22) and (25) concentrate around the values specified in Corollary 1 and Theorem 1, respectively.

Theorem 2. *The loss function of the hybrid JS-estimator in* (20) *satisfies the following:*

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

$$\begin{aligned} \mathbb{P}\Bigg(\Bigg| \frac{1}{n} \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_H}\|^2 - \min\left(\frac{\rho_n \sigma^2}{\rho_n + \sigma^2}, \frac{\beta_n \sigma^2}{g\left(\alpha_n + \sigma^2\right)} \right. \\ \left. + \kappa_n \delta + o(\delta) \right) \Bigg| \geq \epsilon \Bigg) \leq K e^{-\frac{nk \min(\epsilon^2, 1)}{\max(\|\boldsymbol{\theta}\|^2/n, 1)}}, \end{aligned}$$

where ρ_n, α_n, and β_n are respectively given by (19), (16), *and* (17), *and K and k are positive constants.*

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

$$\begin{aligned} \lim_{n\to\infty} \Bigg| \frac{1}{n} R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_H}) - \min\left(\frac{\rho_n \sigma^2}{\rho_n + \sigma^2}, \frac{\beta_n \sigma^2}{g\left(\alpha_n + \sigma^2\right)} \right. \\ \left. + \kappa_n \delta + o(\delta) \right) \Bigg| = 0. \end{aligned}$$

The proof of the theorem in given in [8]. The theorem implies that the hybrid estimator chooses the better of the $\hat{\boldsymbol{\theta}}_{L_+}$ and $\hat{\boldsymbol{\theta}}_{JS_2}$ with high probability, with the probability of choosing the worse estimator decreasing exponentially in n.

IV. Simulation Results

We present simulation plots that compare the average normalized loss of the proposed estimators with that of the regular JS-estimator and Lindley's estimator, for various choices of $\boldsymbol{\theta}$. In each plot, the normalized loss, labelled $\frac{1}{n}\tilde{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ on the Y-axis, is computed by averaging over 1000 realizations of $\mathbf{w}$. We use $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I})$, i.e., the noise variance $\sigma^2 = 1$. Both the regular JS-estimator and Lindley's estimator used are the positive-part versions, respectively given by (6) and (7). We choose $\delta = 5/\sqrt{n}$ for our proposed estimators.

In Fig. 1a and Fig. 1b, the $\{\theta_i\}_{i=1}^n$ are arranged in two clusters; we see that the two-cluster JS-estimator has lower risk than Lindley's estimator because $\frac{\beta_n \sigma^2}{\alpha_n + \sigma^2}$ is smaller than $\frac{\rho_n \sigma^2}{\rho_n + \sigma^2}$ (where $\alpha_n, \beta_n, \rho_n$ are given by (16), (17), (19), respectively). However, the situation is reversed in Fig. 1c where $\{\theta_i\}_{i=1}^n$ are uniformly placed within an interval; we see that Lindley's estimator is the better choice in this case.

The two arrangements of $\{\theta_i\}_{i=1}^n$ are intentionally chosen to highlight the advantage of the hybrid JS-estimator. The plots demonstrate that the hybrid JS-estimator has average loss close to the minimum of those of Lindley's estimator and the two-cluster JS-estimator even for moderately large dimensions, e.g. $n = 50$. This is in agreement with the result of Theorem 2. The losses of the proposed estimators are significantly smaller than that of the regular JS-estimator when the $\{\theta_i\}_{i=1}^n$ are approximately arranged in two separable clusters. More simulation plots that support the claims of Theorems 1 and 2 are provided in [8].

V. Concluding Remarks

In this paper, we presented a two-cluster JS-estimator and its hybrid version that take advantage of the large dimensionality to infer the clustering structure of the parameter values from the observed data. This structure is then used to choose a good attracting vector for the shrinkage estimator. The constructed estimators have significantly smaller risks than the regular (positive-part) JS-estimator for a wide range of $\boldsymbol{\theta} \in \mathbb{R}^n$, even though they do not dominate it for finite n. We obtained concentration bounds for the squared error loss of these estimators, and convergence results for the risk.

In [8], the ideas presented in this paper are further generalized to define and analyze multiple-cluster hybrid JS-estimators. These hybrid estimators consider attracting subspaces with up to L clusters (for any integer $L \geq 2$), and aim to choose the best one for the $\boldsymbol{\theta}$ in context.

Acknowledgement

This work was supported in part by a Marie Curie Career Integration Grant (Grant Agreement No. 631489) and an Early Career Grant from the Isaac Newton Trust.

References

[1] W. James and C. M. Stein, "Estimation with Quadratic Loss," in *Proc. Fourth Berkeley Symp. Math. Stat. Probab.*, pp. 361–380, 1961.

[2] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. Springer, New York, NY, 1998.

[3] B. Efron and C. Morris, "Data Analysis Using Stein's Estimator and Its Generalizations," *J. Amer. Statist. Assoc.*, vol. 70, pp. 311–319, 1975.

[4] D. V. Lindley, "Discussion on Professor Stein's Paper," *J. R. Stat. Soc.*, vol. 24, pp. 285–287, 1962.

[5] B. Efron and C. Morris, "Stein's estimation rule and its competitors—an empirical Bayes approach," *J. Amer. Statist. Assoc.*, vol. 68, pp. 117–130, 1973.

[6] E. George, "Minimax Multiple Shrinkage Estimation," *Ann. Stat.*, vol. 14, pp. 188–205, 1986.

[7] E. George, "Combining Minimax Shrinkage Estimators," *J. Amer. Statist. Assoc.*, vol. 81, pp. 437–445, 1986.

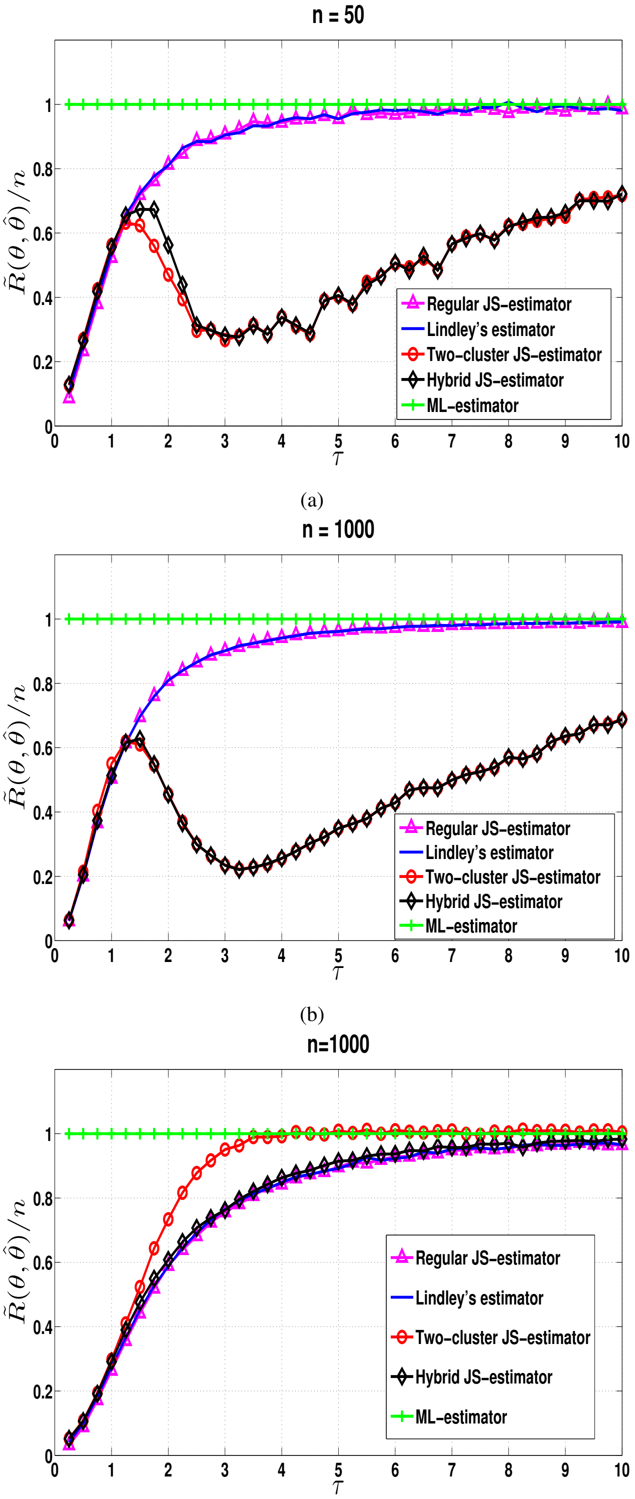
[8] K. P. Srinath and R. Venkataramanan, "Cluster-Seeking James-Stein Estimators," [Online]: http://arxiv.org/abs/1602.00542.

[9] A. J. Baranchik, "Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution," *Tech. Report, 51, Stanford Uni.*, 1964.

[10] P. Shao and W. E. Strawderman, "Improving on the James-Stein Positive Part Estimator," *Ann. Stat.*, vol. 22, no. 3, pp. 1517–1538, 1994.

[11] Y. Maruyama and W. E. Strawderman, "Necessary conditions for dominating the James-Stein estimator," *Ann. Inst. Stat. Math.*, vol. 57, pp. 157–165, 2005.

[12] G. Leung and A. R. Barron, "Information Theory and Mixing Least-Squares Regressions," *IEEE Trans. Inf. Theory*, vol. 52, pp. 3396–3410, Aug. 2006.



(a)

(b)

(c)

Fig. 1. Plots of the average normalized loss of various estimators for different values of n, and for different arrangements of $\{\theta_i\}_{i=1}^n$. In (a), $n = 50$ and in (b), $n = 1000$. In both (a) and (b), $\{\theta_i\}_{i=1}^n$ is divided into two clusters with each cluster having $n/2$ points. The clusters are centred at $-\tau$ and τ and the width of each cluster is chosen to be 0.5τ. The locations of the points within each cluster are chosen uniformly at random. In (c), $n = 1000$, with $\{\theta_i\}_{i=1}^n$ being placed uniformly between $-\tau$ to τ.