Empirical Bayes Estimators for Sparse Sequences

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

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

***Abstract*—The problem of estimating a high-dimensional sparse vector $\boldsymbol{\theta} \in \mathbb{R}^n$ from an observation in i.i.d. Gaussian noise is considered. An empirical Bayes shrinkage estimator, derived using a Bernoulli-Gaussian prior, is analyzed and compared with the well-known soft-thresholding estimator using squared-error loss as a measure of performance. We obtain concentration inequalities for the Stein's unbiased risk estimate and the loss function of both estimators.**

Depending on the underlying $\boldsymbol{\theta}$, either the proposed empirical Bayes (eBayes) estimator or soft-thresholding may have smaller loss. We consider a hybrid estimator that attempts to pick the better of the soft-thresholding estimator and the eBayes estimator by comparing their risk estimates. It is shown that: i) the loss of the hybrid estimator concentrates on the minimum of the losses of the two competing estimators, and ii) the risk of the hybrid estimator is within order $1/\sqrt{n}$ of the minimum of the two risks. Simulation results are provided to support the theoretical results.

I. Introduction

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

$$\mathbf{y} = \boldsymbol{\theta} + \mathbf{w}. \tag{1}$$

The noise vector $\mathbf{w} \in \mathbb{R}^n$ is distributed as $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (by rescaling $\mathbf{y}$ by $1/\sigma$, the case where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ reduces to the above form with $\boldsymbol{\theta}/\sigma$ to be estimated). In this paper, as a measure of the performance of an estimator $\hat{\boldsymbol{\theta}}$, we consider 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:

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

We emphasize that $\boldsymbol{\theta}$ is deterministic, so the expectation above is computed over $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{I})$.

We assume that $\boldsymbol{\theta}$ has k non-zero entries out of n, where k may not be known to the estimator. Though our results are general, they are most interesting for the case where $k = \Theta(n)$. Thus as n gets large, the sparsity level $\eta := k/n$ is bounded above and below by arbitrary constants in $(0, 1]$.

The sparse estimation problem has been widely studied [1]–[7] due to its fundamental role in non-parametric function estimation. If the function has a sparse representation in an orthogonal basis (e.g., a Fourier or wavelet basis), then (1) models the problem of estimating the function from a noisy measurement of n basis coefficients. Another motivation for constructing good sparse estimators comes from Approximate Message Passing (AMP) algorithms for compressed sensing, which is discussed in Sec. V-A.

The soft thresholding estimator is a popular choice of estimator when $\boldsymbol{\theta}$ is assumed to be sparse [1]–[3], [7], [8], and is given as follows for threshold λ. For $i \in \{1, 2, \cdots, n\}$,

$$\hat{\theta}_{ST,i}(y_i; \lambda) = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0, & \text{if } -\lambda \le y_i \le \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda. \end{cases}$$

Along with its simplicity, the soft-thresholding estimator has other attractive properties. For example, when n is large and the sparsity level $\eta = k/n \to 0$, the worst-case risk over the set of η-sparse vectors is $2\eta \log \eta^{-1}(1 + o(1))$. However, no sharp theoretical bounds exist for the risk of the soft-thresholding estimators for moderate or large values of η.

A. Motivation and Contributions

The well-known (positive-part) James-Stein estimator [9] for estimating an arbitrary $\boldsymbol{\theta} \in \mathbb{R}^n$ from an observation $\mathbf{y} = \boldsymbol{\theta} + \mathbf{w}$, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, is given by

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

where X_+ denotes $\max(0, X)$. The James-Stein estimator has uniformly lower risk than the maximum-likelihood estimator $\hat{\boldsymbol{\theta}}(\mathbf{y}) = \mathbf{y}$ (see, e.g., [10, Chap. 5]). An empirical Bayes viewpoint of this estimator is provided in [11]: assuming a Gaussian prior on $\boldsymbol{\theta}$ so that $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \xi^2\mathbf{I})$, the Bayes estimator is

$$\hat{\boldsymbol{\theta}}_{Bayes} = \left(1 - \frac{1}{1+\xi^2}\right)\mathbf{y}. \tag{3}$$

Based on the Gaussian prior, we have $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, (1+\xi^2)\mathbf{I})$. So, $(n-2)/\|\mathbf{y}\|^2$ is an unbiased estimate of $1/(1+\xi^2)$, i.e., $\mathbb{E}\left[(n-2)/\|\mathbf{y}\|^2\right] = 1/(1+\xi^2)$. Plugging in this estimate of $1/(1+\xi^2)$ in (3) (and ensuring that it is always ≤ 1) gives $\hat{\boldsymbol{\theta}}_{JS}$ in (2). One can also start with a Gaussian prior with non-zero mean, i.e., $\boldsymbol{\theta} \sim \mathcal{N}(\mu\mathbf{1}, \xi^2\mathbf{I})$, and use $\sum_i y_i/n$ as a plug-in estimate for μ. The resulting empirical Bayes estimator is the positive-part Lindley's estimator [11].

This empirical Bayes derivation of the James-Stein estimator serves as a motivation for our work. In our setting, since we know that $\boldsymbol{\theta}$ is sparse, we consider an empirical Bayes estimator based on a prior that is a mixture of a point mass at 0 and a continuous distribution with density $\psi(\theta; \mu, \xi)$, where μ is a location parameter (mean) and ξ is a scale parameter. The prior is given by

$$f(\theta; \epsilon, \mu, \xi) = (1-\epsilon)\delta(\theta) + \epsilon\,\psi(\theta; \mu, \xi), \quad \theta \in \mathbb{R}. \tag{4}$$

The parameter $\epsilon \in [0,1]$, which controls the sparsity, is treated as a fixed parameter that can be optimized. In particular, ϵ *need not* be the true sparsity level η (which may be unknown). Taking ψ to be the Gaussian density, in Sec. II we derive an empirical Bayes (eBayes) estimator using plug-in estimates for μ and ξ^2. In Sec. III, we derive a risk function estimate for the eBayes estimator using Stein's unbiased risk estimate (SURE). We then consider a hybrid estimator which chooses between the eBayes estimator and the soft-thresholding estimator by comparing their risk estimates.

Sec. IV contains the main theoretical results of the paper. Theorem 1 shows that for large n, the SURE concentrates on a deterministic value which is within $\mathcal{O}(1/\sqrt{n})$ of the true risk. Theorem 2 shows that the loss of the eBayes estimator concentrates on a deterministic value that is also within $\mathcal{O}(1/\sqrt{n})$ of the risk. Using these results (and analogous ones for soft-thresholding), Theorem 5 shows that for the hybrid estimator, the loss concentrates on the minimum of the losses of the two rival estimators, and its risk is within $\mathcal{O}(1/\sqrt{n})$ of the minimum of the two risks. In Section V, we provide simulation results, including an application of the hybrid estimator in the approximate message passing (AMP) algorithm for compressed sensing.

B. Related Work

In the context of wavelets, several works have considered estimators based on a signal prior that is a mixture of a point mass at 0 and a Gaussian distribution (see, e.g., [12]). In most of these works, the hyperparameters of the prior are chosen based on some prior information about the signal. Empirical Bayes estimators based on a prior that is a mixture of a point mass at 0 and a distribution with a heavy-tailed density have been proposed in [3], [4]. The weights of the mixture are first determined using marginal log-likelihood; the estimator then uses a thresholding rule based on the posterior median. It has been shown that the risk of this estimator over the class of η-sparse vectors is within a constant factor of the minimax risk when the sparsity level η is small enough.

In this paper, we use a fixed mixture weight for the empirical Bayes estimator and empirically estimate the location and scale parameters of the continuous part of the prior. This approach allows us to obtain concentration inequalities for the risk estimates, which then lead to a risk bound for the hybrid estimator.

Notation: The set $\{1, 2, \cdots, n\}$ is denoted by $[n]$. Bold lowercase (uppercase) letters are used to denote vectors (matrices), and plain lowercase letters for their entries. For example, the entries of $\mathbf{y}$ are y_i, $i = 1, \cdots, n$. The indicator function of an event $\mathcal{E}$ is denoted by $1_{\{\mathcal{E}\}}$. For positive-valued functions $f(n)$ and $g(n)$, the notation $f(n) = \mathcal{O}(g(n))$ means that $\exists k > 0$ such that $\forall n > n_0$, $f(n) \leq kg(n)$.

II. Empirical Bayes Estimator

Assuming that $\{\theta_i\}, i \in [n]$, were generated i.i.d. $\sim f$ in (4), the conditional mean of θ given y is the optimal estimator (for squared-error loss). The empirical Bayes estimator for a *fixed* $\epsilon \in [0,1]$ is this conditional mean, with the values of μ, ξ estimated from the data $\mathbf{y}$. Hence, $\forall i \in [n]$,

$$\hat{\boldsymbol{\theta}}_{EB,i}(\mathbf{y};\epsilon) = \frac{\int_{\mathbb{R}} x f(x;\epsilon,\hat{\mu},\hat{\xi})\phi(y_i - x)dx}{\int_{\mathbb{R}} f(x;\epsilon,\hat{\mu},\hat{\xi})\phi(y_i - x)dx}. \tag{5}$$

In (5), $\phi(x) := \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the standard normal density, and $\hat{\mu}, \hat{\xi}$ are the estimates of μ, ξ from $\mathbf{y}$. A consistent estimator for μ (converging in probability to μ) is

$$\hat{\mu}(\mathbf{y}) = \bar{y}/\epsilon, \tag{6}$$

where the empirical mean $\bar{y} := \sum_i y_i/n$. The scale parameter can be estimated using the second moment $\overline{y^2} := \|\mathbf{y}\|^2/n$ and the first moment $\bar{y}$. In this paper, we consider the Gaussian density for ψ so that

$$\psi(\theta;\mu,\xi) = \frac{1}{\sqrt{2\pi\xi^2}} \exp(-(\theta-\mu)^2/2\xi^2).$$

The mean μ is estimated as in (6), and ξ^2, being the variance, is estimated as

$$\widehat{\xi^2}(\mathbf{y}) = \frac{1}{\epsilon}\left(\overline{y^2} - \frac{(\bar{y})^2}{\epsilon} - 1\right)_+.$$

The resulting empirical Bayes estimator is, for $i \in [n]$,

$$\hat{\boldsymbol{\theta}}_{EB,i}(\mathbf{y};\epsilon) = \frac{\hat{\mu} + \left(1 - \frac{1}{1+\widehat{\xi^2}}\right)(y_i - \hat{\mu})}{1 + \frac{(1-\epsilon)}{\epsilon}\sqrt{1+\widehat{\xi^2}}\,\exp\left(\frac{(y_i-\hat{\mu})^2}{2(1+\widehat{\xi^2})} - \frac{y_i^2}{2}\right)}. \tag{7}$$

For $\epsilon = 1$, $\hat{\boldsymbol{\theta}}_{EB}$ reduces to the well-known James-Stein estimator [9], [13] that shrinks each element of $\mathbf{y}$ towards the empirical mean $\bar{y}$.

Note that $\hat{\boldsymbol{\theta}}_{EB}$ is a shrinkage estimator — it shrinks each y_i towards a common element $\hat{\mu}$, with the amount of shrinkage depending on y_i. There are two terms that determine the shrinkage, the first being the term $\left[1 - \frac{1}{1+\widehat{\xi^2}}\right]$ which is common for all the y_i. The second term influencing the shrinkage is the exponential in the denominator which depends on y_i ; the smaller y_i is, the smaller θ_i is expected to be and hence, the larger the amount of shrinkage.

III. Risk Estimators and the Hybrid Estimator

Depending on the underlying $\boldsymbol{\theta}$, either $\hat{\boldsymbol{\theta}}_{ST}$ or $\hat{\boldsymbol{\theta}}_{EB}$ may have smaller loss. To construct a hybrid estimator that reliably chooses the better estimator, we use Stein's unbiased risk estimate (SURE) [14] to estimate the losses of each estimator.

Fact 1: [14] If an estimator $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is almost everywhere differentiable, then the SURE of $\hat{\boldsymbol{\theta}}$, given by

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

is an unbiased estimate of the risk $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$, i.e., $\mathbb{E}[\hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{y}))] = R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$.

Both the risk estimate and the loss function of an estimator $\hat{\boldsymbol{\theta}}$ are random variables depending on $\mathbf{y}$. Henceforth, we do

not explicitly indicate the dependency of the two on $\mathbf{y}$. The normalized SURE for $\hat{\boldsymbol{\theta}}_{ST}$ with threshold λ is given by

$$\frac{\hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ST}; \lambda)}{n} = -1 + \frac{\|\mathbf{y} - \hat{\boldsymbol{\theta}}_{ST}\|^2}{n} + \frac{2}{n}\sum_{i=1}^{n} 1_{\{y_i^2 > \lambda^2\}}. \quad (8)$$

To keep the exposition simple, for our concentration results we assume that the location parameter $\hat{\mu}$ in $\hat{\boldsymbol{\theta}}_{EB}$ is zero. Extending the results to the case with a general $\hat{\mu}$ is straightforward, though a bit cumbersome. Using SURE, the normalized risk estimate for $\hat{\boldsymbol{\theta}}_{EB}$ with $\hat{\mu} = 0$ is

$$\begin{aligned}\frac{\hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB}; \epsilon)}{n} &= \left(\frac{\|\mathbf{y}\|^2}{n} - 1\right) \\ &+ \frac{a_{\mathbf{y}}^2}{n}\sum_{i=1}^{n} \frac{y_i^2\left(1 + 2c_{\mathbf{y}} e^{-\frac{a_{\mathbf{y}} y_i^2}{2}}\right)}{b_i^2(\mathbf{y})} - \frac{2a_{\mathbf{y}}}{n}\sum_{i=1}^{n}\frac{y_i^2 - 1}{b_i(\mathbf{y})} \\ &+ \frac{4}{d_{\mathbf{y}}^2 \epsilon n^2}\sum_{i=1}^{n}\frac{y_i^2}{b_i(\mathbf{y})} 1_{\{\|\mathbf{y}\|^2 > n\}} + \frac{2(1-\epsilon)a_{\mathbf{y}}}{d_{\mathbf{y}}^{3/2}\epsilon^2 n^2}\sum_{i=1}^{n}\frac{y_i^4 e^{-\frac{a_{\mathbf{y}} y_i^2}{2}}}{b_i^2(\mathbf{y})} \\ &- \frac{2(1-\epsilon)a_{\mathbf{y}}}{\sqrt{d_{\mathbf{y}}}\epsilon^2 n^2}\sum_{i=1}^{n}\frac{y_i^2 e^{-\frac{a_{\mathbf{y}} y_i^2}{2}}}{b_i^2(\mathbf{y})} \end{aligned} \quad (9)$$

where

$$\begin{aligned} a_{\mathbf{y}} &:= \frac{\widehat{\xi}^2}{1 + \widehat{\xi}^2} = \left[1 - \frac{\epsilon}{(\|\mathbf{y}\|^2/n - 1)_+ + \epsilon}\right], \\ d_{\mathbf{y}} &:= 1 + \widehat{\xi}^2 = 1 + \frac{1}{\epsilon}\left(\frac{\|\mathbf{y}\|^2}{n} - 1\right)_+, \\ c_{\mathbf{y}} &:= \frac{1-\epsilon}{\epsilon}\sqrt{1 + \widehat{\xi}^2} = \frac{1-\epsilon}{\epsilon}\sqrt{d_{\mathbf{y}}}, \\ b_i(\mathbf{y}) &:= 1 + c_{\mathbf{y}} e^{-\frac{a_{\mathbf{y}} y_i^2}{2}}. \end{aligned}$$

For large n, it is shown in [15, Lemma 4.1] that the last three terms in (9) each concentrate around deterministic constants of order $1/n$. These terms can therefore be neglected in a practical application of the risk estimate.

We use the risk estimates in (8) and (9) to define a hybrid estimator that aims to select the estimator with smaller loss for the $\boldsymbol{\theta}$ in context. The hybrid estimator is defined as

$$\hat{\boldsymbol{\theta}}_H = \gamma_{\mathbf{y}}\hat{\boldsymbol{\theta}}_{EB} + (1 - \gamma_{\mathbf{y}})\hat{\boldsymbol{\theta}}_{ST}, \quad (10)$$

where

$$\gamma_{\mathbf{y}} = \begin{cases} 1 & \text{if } \hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB}) \le \hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ST}), \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

In the next section, we present concentration results for the risk estimates and loss functions of $\hat{\boldsymbol{\theta}}_{ST}$ and $\hat{\boldsymbol{\theta}}_{EB}$, and use these to show that the loss of the hybrid estimator concentrates on the minimum of the losses of the two estimators. Due to space constraints, we omit the proofs, which can be found in Sections 4 and 6 of [15].

IV. Main Results

The constants in our concentration results for the eBayes estimator depend on $\boldsymbol{\theta}$ via $\frac{1}{n}\sum_{i=1}^{n}\theta_i^4$. In order to make these constants universal, we assume that the fourth moment of $\boldsymbol{\theta}$ is bounded.

Assumption A: There exists a finite constant $\Lambda > 0$ such that $\frac{1}{n}\sum_{i=1}^{n}\theta_i^4 \le \Lambda$.

When Assumption A is satisfied, the constants in the concentration results depend only on Λ (and not on the underlying $\boldsymbol{\theta}$ or n). For brevity, we henceforth do not explicitly indicate the dependence on λ and ϵ in the notation for the risk estimates on the LHS of (8) and (9), respectively.

Theorem 1: Consider a sequence of $\boldsymbol{\theta}$ with increasing dimension n and satisfying Assumption A. Then the risk estimate $\hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB})$ satisfies the following for any $t > 0$:

$$\mathbb{P}\left(\frac{1}{n}\left|\hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB}) - R_1(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB})\right| \ge t\right) \le Ke^{-nk\min(t, t^2)}$$

where $0 < K \le 24$ and $k > 0$ are absolute constants, and $R_1(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB})$ is a deterministic quantity such that

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

The next result shows that, like the risk estimate, the normalized loss of the eBayes estimator also concentrates on a deterministic value close to the true risk.

Theorem 2: Consider a sequence of $\boldsymbol{\theta}$ with increasing dimension n and satisfying Assumption A. Then the loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB}) = \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{EB}\|^2$ satisfies the following for any $t > 0$:

$$\mathbb{P}\left(\frac{1}{n}\left|L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB}) - R_2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB})\right| \ge t\right) \le Ke^{-nk\min(t, t^2)}$$

where $K \le 10$ and k are absolute positive constants, and $R_2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB})$ is a deterministic quantity such that

$$\left|\frac{R_2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB})}{n} - \frac{R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB})}{n}\right| = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

The normalized SURE and the normalized loss for $\hat{\boldsymbol{\theta}}_{ST}$ with threshold λ satisfy the following:

Theorem 3: [8] The SURE $\hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ST}; \lambda)$ for the soft-thresholding estimator with parameter λ satisfies, for any $t > 0$,

$$\mathbb{P}\left(\frac{1}{n}\left|\hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ST}) - R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ST})\right| \ge t\right) \le 2e^{-\frac{2t^2}{9(1+\lambda^2)^2}}.$$

Theorem 4: The loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ST}) = \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{ST}\|^2$ of the soft-thresholding estimator satisfies the following for any $t > 0$:

$$\mathbb{P}\left(\frac{1}{n}\left|L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ST}) - R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ST})\right| \ge t\right) \le 2e^{-nk\min(t, t^2)}$$

where k is an absolute positive constant.

For a given $\boldsymbol{\theta}$, let

$$L_{min}(\boldsymbol{\theta}, \mathbf{y}) := \min\left\{L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB}), L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ST})\right\},$$
$$\kappa_n := \frac{1}{n}\left|R_1(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB}) - R_2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB})\right|,$$

where $R_1(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB})$ and $R_2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB})$ are the deterministic concentrating values in Theorems 1 and 2, respectively. Note that κ_n is an $\mathcal{O}(1/\sqrt{n})$ quantity since both $R_1(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB})/n$ and $R_2(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB})/n$ are within $\mathcal{O}(1/\sqrt{n})$ from $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB})/n$. The following theorem characterizes the loss $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_H(\mathbf{y}))$ and the risk $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_H)$ of the hybrid estimator.

Theorem 5: Consider a sequence of $\boldsymbol{\theta}$ with increasing dimension n and satisfying Assumption A. Then, for any $t > 0$, we have

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

for some absolute positive constants K and k. The risk of the hybrid estimator can be bounded as

$$\frac{R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_H)}{n} \leq \frac{1}{n}\min\left\{R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB}),\, R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ST})\right\} + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

V. Simulation Results

When the true sparsity level η is unknown, one can optimize SURE to find the best fit for both $\hat{\boldsymbol{\theta}}_{ST}$ and $\hat{\boldsymbol{\theta}}_{EB}$. The concentration results (Theorems 1 and 3) imply that the SURE for either estimator does not deviate much from the actual risk for large n. SureShrink, proposed in [8], chooses the thresholding parameter λ^* as follows, from a set $\mathcal{S}$ that is a discretized version of the interval $(0, \sqrt{2\log n}]$.

$$\lambda^* = \arg\min_{\lambda \in \mathcal{S}} \hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ST}; \lambda)/n \tag{12}$$

where $\hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ST}; \lambda)$ is defined in (8).

We propose to find the best value of ϵ in (7) by first discretizing the set $(0, 1]$ (denoting it by $\mathcal{D}$), and choosing the sparsity parameter as

$$\epsilon^* = \arg\min_{\epsilon \in \mathcal{D}} \hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB}; \epsilon)/n. \tag{13}$$

Here $\hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB}; \epsilon)/n$ is as in (9), with suitable modifications to account for non-zero $\hat{\mu}$. The hybrid estimator then chooses the estimator with lower value of SURE ($\hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ST}; \lambda^*)$ vs. $\hat{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{EB}; \epsilon^*)$).

Fig. 1 shows the average normalized loss $\tilde{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})/n$ (averaged over 100 realizations of $\mathbf{w}$) of the three estimators at different sparsity levels for two choices of the distribution of the non-zero entries of $\boldsymbol{\theta}$. We assume that the actual sparsity factor η is unknown and use SURE to find the best sparsity parameters for $\hat{\boldsymbol{\theta}}_{ST}$ and $\hat{\boldsymbol{\theta}}_{EB}$. The optimization is performed over the discrete sets $\mathcal{D} = \{0.02i, i \in [50]\}$ and $\mathcal{S} = \{0.1i, i \in [\lceil 10\sqrt{2\log n}\rceil]\}$. In all the plots, $n = 1000$. Additional simulation plots, including for n as low as 50, are provided in [15, Section 5]. The plots suggest that for a wide range of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}_{EB}$ is at least as good as $\hat{\boldsymbol{\theta}}_{ST}$ for all values of the sparsity factor η, and better in most cases.

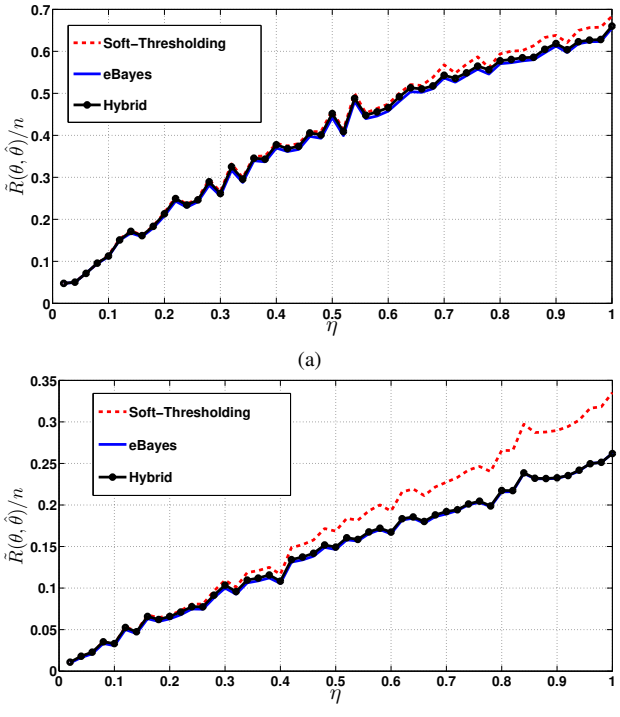

Fig. 1. Average normalized loss $\tilde{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})/n$ with $n = 1000$ for the following cases: a) The non-zero entries are drawn from the Laplace distribution with mean 0 and variance 2. b) The non-zero entries are drawn from the uniform distribution on the interval $[-2, 2]$.

A. Application to Compressed Sensing

In compressed sensing, the goal is to estimate a sparse vector $\boldsymbol{\theta} \in \mathbb{R}^n$ from a noisy linear measurement $\mathbf{y} \in \mathbb{R}^m$ of the form

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \mathbf{w}.$$

Assume that $\mathbf{A}$ is an $m \times n$ random matrix with i.i.d. sub-Gaussian entries with variance $1/m$, and the noise vector $\mathbf{w} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. The undersampling ratio is denoted by $\delta := m/n < 1$.

For this model, Approximate Message Passing (AMP) [16]–[18] is a class of iterative algorithms to estimate $\boldsymbol{\theta}$ from $\mathbf{y}$. Starting with the initial conditions $\boldsymbol{\theta}_0 = 0$, $\mathbf{z}_0 = \mathbf{y}$, AMP iteratively produces estimates $\{\boldsymbol{\theta}_{t+1}\}_{t\geq 0}$ as follows [17]:

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= f_t\left(\mathbf{A}^T\mathbf{z}_t + \boldsymbol{\theta}_t\right) \\ \mathbf{z}_{t+1} &= \mathbf{y} - \mathbf{A}\boldsymbol{\theta}_{t+1} + \frac{1}{\delta}\mathbf{z}_t\left\langle f_t'\left(\mathbf{A}^T\mathbf{z}_t + \boldsymbol{\theta}_t\right)\right\rangle.\end{aligned} \tag{14}$$

Here for each t, $f_t : \mathbb{R} \to \mathbb{R}$ is a "denoising" function, f_t' denotes its derivative, and both functions act component-wise on vectors. For $\mathbf{u} \in \mathbb{R}^n$, $\langle \mathbf{u} \rangle$ denotes the average of its entries.

The AMP update (14) is underpinned by the following key property of the effective observation vector $(\mathbf{A}^T\mathbf{z}_t + \boldsymbol{\theta}_t)$: for large n, after each iteration t, $(\mathbf{A}^T\mathbf{z}_t + \boldsymbol{\theta}_t)$ is approximately distributed as $\boldsymbol{\theta} + \tau_t\mathbf{Z}$, where $\mathbf{Z} \in \mathbb{R}^n$ is a standard Gaussian vector that is independent of $\boldsymbol{\theta}$. The effective noise variance τ_t^2 is determined (in the large system limit) by a scalar recursion

called state evolution [17]. For our purposes, it suffices to note that a good estimate of τ_t^2 is given by $\widehat{\tau}_t^2 := \|\mathbf{z}_t{}^2\|/m$.

The function f_t estimates the sparse vector $\boldsymbol{\theta}$ from an observation in Gaussian noise of variance approximately $\widehat{\tau}_t^2$. Therefore, in each iteration, the AMP provides a platform to compare the performance of soft-thresholding and the eBayes estimator (and hence the hybrid estimator) as choices for f_t. We note that while soft-thresholding operates on a vector component-wise, the eBayes estimator doesn't. However, for sufficiently large values of m and n, both $\hat{\mu}$ and $\widehat{\xi^2}$ in (6)-(7) are close to deterministic values in which case the eBayes estimator also approximately acts component-wise on a vector.

The simulation plots in Fig. 2 show the performances of the three estimators when used in the AMP algorithm. We fix $n = 10000$ and consider two set-ups, which differ in the undersampling ratio $\delta = m/n$, sparsity factor $\eta = \|\boldsymbol{\theta}\|_0/n$, noise variance σ^2, and the non-zero values of $\boldsymbol{\theta}$. The measurement matrix $\mathbf{A}$ is chosen with its entries i.i.d. $\sim \mathcal{N}(0, 1/m)$, and the sparsity factor η is assumed to be unknown. So, at each step of the algorithm, a suitable threshold λ_i^* (for soft-thresholding) and a suitable sparsity parameter ϵ_i^* (for the eBayes estimator) are chosen as described in (12) and (13) with the only difference being that the optimization is now based on $\|\mathbf{z}_t\|^2/n$, and not on SURE. A precise description of the algorithm can be found in [15, Section 5.1].

The plots in Fig. 2 show the progression of the mean squared error (MSE) $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}\|^2/n$ with the AMP iteration number t for the three estimators when applied in the AMP algorithm for compressed sensing. It can be inferred that the eBayes estimator provides a strong alternative to soft-thresholding in the AMP.

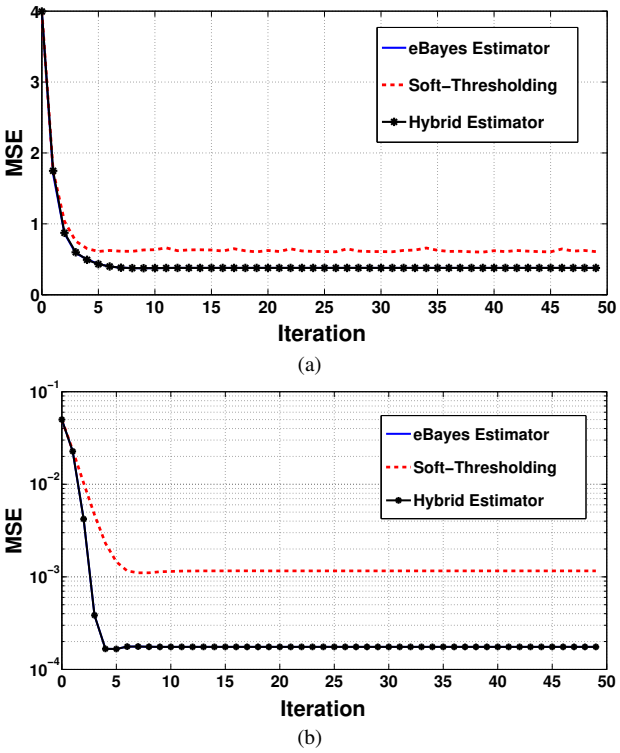

Fig. 2. Plots of the mean squared error $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}\|^2/n$ as a function of the iteration number t for the following cases: a) $\delta = 0.65$, $\epsilon = 0.13$, $\sigma = 1$, the non-zero entries of $\boldsymbol{\theta}$ are drawn from $\mathcal{N}(0, 5)$. b) $\delta = 0.5$, $\epsilon = 0.05$, $\sigma = 0.05$, the non-zero entries are drawn from the Rademacher distribution.

VI. Concluding Remarks

For the problem of estimating a sparse vector (with possibly unknown sparsity level), we proposed an empirical Bayes estimator based on a Bernoulli-Gaussian prior. By obtaining a concentration inequality for its risk estimate (SURE), we showed that the risk of the hybrid estimator is close to the minimum of the risks of the competing estimators.

More generally, the approach of Theorem 5 could be used to bound the risk of a hybrid estimator that picks one among several estimators, provided one has concentration bounds for the risk estimates of each of the competing estimators. This suggests that an interesting direction for research is to obtain concentration bounds for the risk estimates of other useful estimators whose parameters depend on the data, e.g., an empirical Bayes estimator based on a Bernoulli-Laplace prior.

References

[1] D. L. Donoho and I. M. Johnstone, "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[2] D. L. Donoho and I. M. Johnstone, "Minimax risk over l_p-balls for l_q-error," *Probab. Th. Rel. Fields*, vol. 99, pp. 277–303, 1994.

[3] I. M. Johnstone and B. W. Silverman, "Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences," *Ann. Stat.*, vol. 32, no. 4, pp. 1594–1649, 2004.

[4] I. M. Johnstone and B. W. Silverman, "Empirical Bayes selection of wavelet thresholds," *Ann. Stat.*, vol. 33, no. 4, pp. 1700–1752, 2005.

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

[6] C. Carvalho, N. Polson, and J. G. Scott, "The horseshoe estimator for sparse signals," *Biometrika*, vol. 97, no. 2, pp. 465–480, 2010.

[7] I. M. Johnstone, *Gaussian estimation: Sequence and wavelet models*. Monograph, Available [Online]: http://statweb.stanford.edu/~imj/GE09-08-15.pdf, 2015.

[8] D. L. Donoho and I. M. Johnstone, "Adapting to Unknown Smoothness via Wavelet Shrinkage," *J. Amer. Stat. Assoc.*, vol. 90, pp. 1200–1224, Dec. 1995.

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

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

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

[12] F. Abramovich, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 60, no. 4, pp. 725–749, 1998.

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

[14] C. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Stat.*, vol. 9, pp. 1135–1151, 1981.

[15] K. P. Srinath and R. Venkataramanan, "Empirical bayes estimators for high-dimensional sparse vectors," *https://arxiv.org/abs/1707.09161*, 2017.

[16] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18914–18919, 2009.

[17] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, 2011.

[18] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory*, pp. 2168–2172, 2011.