Outlier Detection in Large-Scale Sensor Network Data Using Shrinkage Estimators

Ming-Chun Wu and Kwang-Cheng Chen,
Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan
Email : r02942050@ntu.edu.tw and ckc@ntu.edu.tw

***Abstract*—Outlier detection is a frequently encountered technology challenge for many diverse applications in sensor networks, and remains an open problem in general. There are two major difficulties of developing outlier detectors with sensor data. One is the inevitable multi-source identification, the other is the effective inference when discovering information from unknown structured large-scale data. It is even more interesting and challenging with limited observations, since conventional data analysis requires many samples to achieve a satisfactory performance. In this paper, we systematically develop effective and efficient outlier identifiers in parametric and non-parametric ways using shrinkage methodology, like the James-Stein estimator, as the post-processor. We show the superiority of our approach, particularly for the large-scale situations. We further supply a water-filling type algorithm to obtain the asymptotic optimal method for a general class of shrinkage estimators, for wide applications of data analysis.**

I. Introduction

Outlier detection is an important and inevitable problem on account of its influential and diverse applications in many areas. It is frequently encountered particularly in sensor networks like environmental monitoring. In modern large-scale sensor networks, the expectation of detecting multiple outliers from a large number of potential sources makes the problem even more interesting and challenging.

One technology challenge of outlier detection in sensor networks is the multi-source identification, which is to extract useful data from corrupted observations due to multi-source interference. The complexity of optimal approach growths exponentially as the increase of the number of potential outliers making multi-source identification an challenging problem.

The other challenge is to discover the anomaly or unexpected behaviors from the extracted data via data analysis. The performance of conventional data analysis depends on the sample size, and hence we require sufficient samples to achieve satisfactory results. However, even under big data scenario with high-dimensional observations, there might exist low-dimensional information-bearing data that is effective to outlier detections. It suggests the possibility of very limited number of useful data or very limited observation depth, to form a limited-sample detection, particularly outlier detections. As a consequence, due to the inefficiency of conventional data analysis, we have to develop methodology of outlier detection for large-scale sensor data but actually of limited number of information-bearing data.

In this paper, we focus on information collection sensor networks, like environmental monitoring application, and systematically propose outlier identifiers under different levels of prior information of outliers. The structure presented in this paper is general to serve the foundation of solving a class of outlier detections for sensed data. Our methodology consists of two parts. We adopt the subspace method for multi-source identification, due to its feasible complexity along with acceptable performance. We then use shrinkage estimators in proposed identifiers such that they are effective and efficient even in large-scale but limited-sample situations, and propose an algorithm to obtain asymptotic optimal shrinkage estimators for general data analysis.

Shrinkage estimators play a decisive role in large-scale data analysis, and the origin can be traced back to the celebrated *James-Stein estimator* (JS), which shows the superiority of shrinkage estimators over unbiased estimators in the mean estimation of multivariate normal. Inspired by JS, many excellent works from the empirical Bayesian viewpoint have shown the success of shrinkage estimators in statistical data analysis. Moreover, advanced theory developed recently shows that the asymptotic shrinkage estimator can be obtained by solving a constrained optimization, and the solution can be obtained by the proposed water-filling type algorithm in this paper.

A comprehensive survey of outlier detection in sensor networks is presented in [1], while [2] and [3] consider general large-scale outlier detections from the viewpoint of data analysis. Classic works of multi-source identification using the subspace method are the MUSIC algorithm in array signal processing [4] and blind multi-user detection in [5], [6]. In large scale inference, the brilliant James-Stein estimator proposed in [7] along with empirical Bayesian data analysis [8]–[10] are the fundamentals of shrinkage estimators. Moreover, theory of asymptotic optimality in [11] leads to a unified framework of general shrinkage estimators.

The structure of this paper is as follows. We use a general model of the sensor data and organize our approach into three steps. First, we use the subspace method to extract useful data from raw observation in the preprocessing step. Second, we apply conventional statistical inference to identify outliers from extracted data, and develop the main structures of outlier detectors in parametric and nonparametric ways. Third, using shrinkage estimators as post-processor, we develop the outlier detectors which are effective in large-scale situations. At the end, simulations show the superiority of our method in large-scale situations particularly with limited samples.

978-1-4799-5952-5/15/$31.00 ©2015 IEEE

II. Problem Formulation

The observed data collected by M sensors in two-dimensional space with location vectors $\{\boldsymbol{u}_{1:M}\}$ are the measurements taken from a particular physical filed. Meanwhile, K sources with location vectors $\{\boldsymbol{v}_{1:K}\}$ appear and introduce variations to the filed, and thus $\{\boldsymbol{v}_{1:K}\}$ can be estimated by the measurements of the field. Assume the aggregate field is the linear combination of individual field of each source, the m-th sensor measurement at n-th sample is

$$y_m[n] = \sum_{k=1}^{K} f(\boldsymbol{u}_m, \boldsymbol{v}_k) x_k[n] + w_m[n]. \tag{1}$$

The scalar function f model the spatial distribution of the field, $x_k[n]$ is the signal of k-th source at n-th sample and $w_m[n]$ is the additive noise. Define $M \times K$ matrix $\mathbf{H}$ with $\mathbf{H}_{mk} = f(\boldsymbol{u}_m, \boldsymbol{v}_k)$ and collect N samples, we can express the observations in vector form,

$$\boldsymbol{y}[n] = \mathbf{H}\boldsymbol{x}[n] + \boldsymbol{w}[n], \quad n = 1, 2, \ldots, N. \tag{2}$$

In general, measurement noise $w_m[n]$ is assumed to be zero mean white Gaussian noise such that $W_m[n] \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2), \forall m, n$, and signal $x_k[n]$ is modelled as random process due to the incomplete knowledge of signals in general outlier detection. Without loss of generality, we assume $X_k[n] \overset{indep.}{\sim} X_k$ for all k, n. In many scenarios like environmental monitoring, source signals follow right-tailed/nonnegative distributions and sources with strong signal are considered as outliers. Hence, source k is considered as outlying source if it has high probability of having large X_k.

Definition 1. *Outlying Source*
Given $\pi \in [0, 1]$ and outlier threshold $t_{\text{out}} \in \mathbb{R}$. The tail probability of X_k respect to t_{out} is $\pi_k(t) = Pr\{X_k > t_{\text{out}}\}$. Source k is an level π outlying source if $\pi_k(t_{\text{out}}) \geq \pi$.

Tail probability of X_k is an important statistic of source k, and can be estimated without the probability distributions of X_k. In some cases, we know X_k follows a parametric family of probability distributions, and hence tail probabilities can be estimated in a parametric way. Without loss of generality, assume log-normal signal model due to its wide usage in many applications like environmental monitoring, we have $X_k \sim \text{logN}(\mu_k, \sigma_k^2)$ with unknown μ_k and σ_k^2. Then, both parametric and nonparametric approaches are developed in this paper.

From our formulation, one can clearly recognize the two major challenges of developing outlier detectors: 1) multi-source identification: to recover $\boldsymbol{x}[n]$ from $\boldsymbol{y}[n]$, 2) large-scale inference: to estimate tail probabilities simultaneously when K is large. Hence, we systematically and simultaneously solve these two problems in the following sections.

III. Preprocessor

The first step is to extract $\boldsymbol{x}[n]$ from $\boldsymbol{y}[n]$, we thus require a preprocessor to null out the effect of matrix $\mathbf{H}$ and noise $\boldsymbol{w}[n]$ simultaneously. In communication theory, $\mathbf{H}$ is called channel matrix which represents the effect of interference and $\boldsymbol{x}[n]$ can be recovered after equalization. In practice, f is usually known due to our domain knowledge, and sensor location $\boldsymbol{u}_m$ is provided by sensor networks. However, source number K and locations $\{\boldsymbol{v}_{1:K}\}$ are unknown, we apply the subspace method to estimate $\mathbf{H}$ as in [4]–[6] due to its feasible complexity and acceptable performance.

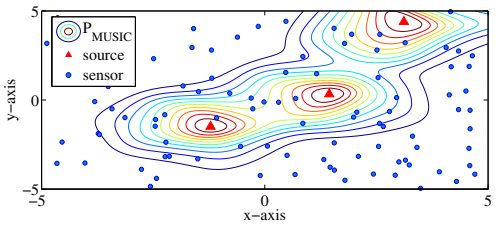

Fig. 1. A realization of $\text{P}_{\text{MUSIC}}(\boldsymbol{v})$ with $f(\boldsymbol{u}, \boldsymbol{v}) = \exp(-\frac{1}{2}\|\boldsymbol{u} - \boldsymbol{v}\|^2)$.

A. Blind Channel Estimation

To have identifiable condition, we assume $\mathbf{H}$ has unknown full rank $K \leq M$ and apply the MUSIC algorithm to estimate both K and $\mathbf{H}$. Note that the covariance matrix of observed data is $\mathbf{C}_{\boldsymbol{yy}} = \mathbf{H}\mathbf{C}_{\boldsymbol{xx}}\mathbf{H}^T + \sigma_w^2\mathbf{I}$. In general, covariance matrix of $\boldsymbol{x}$, $\mathbf{C}_{\boldsymbol{xx}}$, is nonsingular such that $\mathbf{C}_{\boldsymbol{yy}}$ has eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_{K+1} = \cdots = \lambda_M = \sigma_w^2$. Moreover, the signal subspace $\mathbf{E}_s$, the column space of $\mathbf{H}$, is spanned by the eigenvectors corresponding to the K largest eigenvalues. Note that all columns of $\mathbf{H}$ has the form $\boldsymbol{h}(\boldsymbol{v}) = [f(\boldsymbol{u}_1, \boldsymbol{v})\ f(\boldsymbol{u}_2, \boldsymbol{v}) \cdots f(\boldsymbol{u}_M, \boldsymbol{v})]^T$. If there is an source at $\boldsymbol{v}$, $\boldsymbol{h}(\boldsymbol{v})$ must lie in the signal subspace such that

$$\|\mathbf{E}_s^T \boldsymbol{h}(\boldsymbol{v})\|_2^2 / \|\boldsymbol{h}(\boldsymbol{v})\|_2^2 = 1. \tag{3}$$

In practice, we apply eigenvalue decomposition to the sample covariance matrix $\hat{\mathbf{C}}_{\boldsymbol{yy}}$. Then K is determined such that the ratio $\sum_{i=1}^{K} \hat{\lambda}_i / \sum_{i=1}^{M} \hat{\lambda}_i$ first exceeds a predetermined threshold less than but approximating unity. There are some information theoretic methods estimating K without subjective thresholds [12]. Once K is estimated the signal subspace is determined by $\hat{\mathbf{E}}_s = [\hat{\boldsymbol{e}}_1\ \hat{\boldsymbol{e}}_1\ \cdots \hat{\boldsymbol{e}}_K]$ where $\hat{\boldsymbol{e}}_i$ is the eigenvector of $\hat{\mathbf{C}}_{\boldsymbol{yy}}$ with respect to the i-th largest eigenvalue $\hat{\lambda}_i$. Then, define the real function

$$\text{P}_{\text{MUSIC}}(\boldsymbol{v}) = \|\hat{\mathbf{E}}_s^T \boldsymbol{h}(\boldsymbol{v})\|_2^2 / \|\boldsymbol{h}(\boldsymbol{v})\|_2^2, \tag{4}$$

$\{\boldsymbol{v}_{1:K}\}$ are determined by the values of $\boldsymbol{v}$ correspond to the largest K peaks of $\text{P}_{\text{MUSIC}}(\boldsymbol{v})$ as shown in Fig. 1. After all, we have the identify the channel matrix $\mathbf{H}$ as well as source locations $\{\boldsymbol{v}_{1:K}\}$.

B. Equalization

Once we estimate $\mathbf{H}$ from the MUSIC algorithm, equalization can be applied to recover $\boldsymbol{x}[n]$ as in communication theory. The maximum likelihood and least square methods are equivalent when noise is AWGN. To recover the nonnegative signal $\boldsymbol{x}[n]$ from $\boldsymbol{y}[n]$, we have the constrained optimization

$$\underset{\boldsymbol{x} \geq \mathbf{0}}{\text{minimize}} \quad \|\boldsymbol{y}[n] - \mathbf{H}\boldsymbol{x}\|_2^2. \tag{5}$$

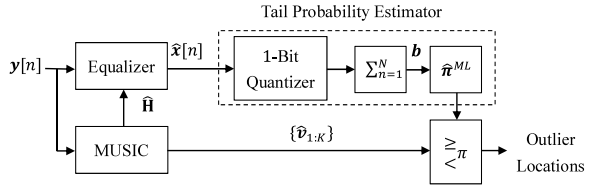

Fig. 2. Nonparametric identifier with ML tail probability estimator (9).

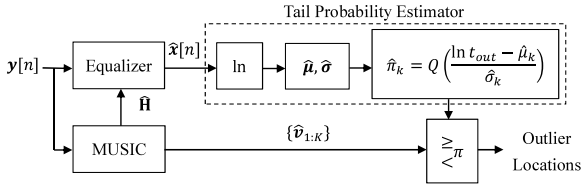

Fig. 3. Block diagram of parametric outlier identifier using (10) and (11).

It is reasonable to have more sensors than sources, that is $M \geq K$, $\mathbf{H}$ thus has linear independent columns. Then, the meaningful solution of (5) is used to identify outliers via data analysis in the following sections.

IV. Data Analysis

To identify outliers we have the K hypothesis testings,

$$\begin{array}{ll} H_{0k}: \text{Source } k \text{ is not outlier}, & \pi_k(t_{\text{out}}) < \pi \\ H_{1k}: \text{Source } k \text{ is outlier}, & \pi_k(t_{\text{out}}) \geq \pi \end{array} \tag{6}$$

Replacing the unknown $\pi_k(t_{\text{out}})$ by its estimation, the decision rules are

$$\hat{\pi}_k(t_{\text{out}}) \underset{H_{0k}}{\overset{H_{1k}}{\gtrless}} \pi, \quad k = 1, 2, \ldots, K, \tag{7}$$

which requires accurate estimation of $\boldsymbol{\pi}(t_{\text{out}})$. Hence we start to estimate $\boldsymbol{\pi}(t_{\text{out}})$ under both parametric and nonparametric cases in this section. For simplicity, we omit t_{out} in the notations of tail probabilities and use $\boldsymbol{x}[n]$ instead of $\hat{\boldsymbol{x}}[n]$.

A. Nonparametric Identifier

First consider the nonparametric case, note that the sufficient statistic of π_k is $b_k = \sum_{n=1}^{N} 1_{(t_{\text{out}},\infty)}(x_k[n])$, and we have

$$B_k | \boldsymbol{\pi} \overset{\text{indep.}}{\sim} \text{Binomial}(N, \pi_k), \ \forall k. \tag{8}$$

The *Uniform Minimum Variance Unbiased Estimator* (UMVUE) is the ML estimator,

$$\hat{\pi}_k^{ML} = \frac{b_k}{N}, \tag{9}$$

since b_k is sufficient and complete statistic of π_k [13]. Due to the superiority among the class of unbiased estimators, we adopt ML estimator and construct the structure of nonparametric identifier as in Fig. 2.

B. Parametric Identifier

In parametric case, we know $X_k[n] \overset{\text{indep.}}{\sim} \log\text{N}(\mu_k, \sigma_k^2)$, $\forall n, k$. To estimate μ_k and σ_k^2, the conventional approach leads to the sample mean and sample variance of the log-transformed signals,

$$\hat{\mu}_k = \frac{1}{N} \sum_{n=1}^{N} \ln(x_k[n]), \tag{10}$$

$$\hat{\sigma}_k^2 = \frac{1}{N-1} \sum_{n=1}^{N} (\ln(x_k[n]) - \hat{\mu}_k)^2. \tag{11}$$

Once $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ are determined we can estimate π_k with the Q-function, $\hat{\pi}_k = Q(\frac{\ln t_{\text{out}} - \hat{\mu}_k}{\hat{\sigma}_k})$. Again, both $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ are UMVUE [13], we use (10) and (11) to construct the structure of parametric identifier as in Fig. 3.

V. James-Stein Estimator

The estimators (9), (10) and (11) are not effective when N is small, the total estimation error thus degrades dramatically as K increases. Moreover, it seems impossible to improve the systems in Fig. 2 and 3, since all estimators used are UMVUE. However, James and Stein show the superiority of biased estimators over UMVUE when estimating mean of multivariate normal distribution in large-scale situations [7]. Actually, both parametric and nonparametric tail probability estimations can be transformed to the mean estimation of multivariate normal. Hence, with the aid of James-Stein (JS) estimator we can improve the identifiers in Fig. 2 and 3.

For nonparametric case, apply the *variance-stabilizing transform* (VST) on equation (8), we have

$$Z_k = \arcsin \sqrt{\frac{B_k + 1/4}{N + 1/2}}. \tag{12}$$

The transformed data has equal variance and good normality [14], that is

$$Z_k \dot{\sim} N(\theta_k, \frac{1}{4N}), \quad \theta_k = \arcsin \sqrt{\pi_k}. \tag{13}$$

For nonparametric case, we have a similar situation after unit variance scaling of the log-transformed signal. Now, the problem is equivalent to estimate $\boldsymbol{\theta}$ with $Z_k \overset{\text{indep.}}{\sim} N(\theta_k, A)$. Under the reasonable normalized squared error loss, $l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{K}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2$, and corresponding risk, $\mathcal{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = E[l(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})]$, there exists a biased estimator outperforming the ML method, $\hat{\boldsymbol{\theta}}^{ML} = \boldsymbol{z}$, which is UMVUE.

Theorem 1. *The James-Stein Estimator [7]*
Let $Z_k \overset{indep.}{\sim} N(\theta_k, A)$, $\bar{z} = \frac{1}{N}\sum_{k=1}^{K} z_k$ *and* $\mathcal{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = E[\frac{1}{K}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2]$. *The James-Stein estimator of* θ_k *is*

$$\hat{\theta}_k^{JS} = (1 - \rho^{JS}) z_k + \rho \bar{z}, \tag{14}$$

$$\rho^{JS} = \min\left(1, \frac{(K-3)A}{\sum_{k=1}^{K}(z_k - \bar{z})^2}\right), \tag{15}$$

and $\mathcal{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{JS}) \leq \mathcal{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{ML}), \ \forall \boldsymbol{\theta} \in \mathbb{R}^K$ *if* $K \geq 4$.

The result is developed by following the philosophy of frequentist statistics, hence JS estimator is *uniformly* better

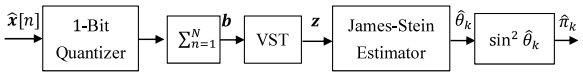

Fig. 4. Nonparametric tail probability estimator using James-Stein estimator.

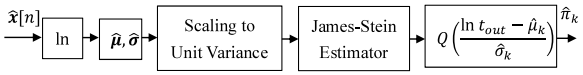

Fig. 5. Parametric tail probability estimator using James-Stein estimator.

Fig. 6. Parametric tail probability estimator using SURE shrinkage estimator.

than ML no matter what $\boldsymbol{\theta}$ is. To get a quick understanding and deep insight from (14), consider the hypothesis testing,

$$H_0 : \text{all } \theta_k\text{'s are equal}, \quad H_1 : \text{not } H_0. \tag{16}$$

If A is known, the F-statistic for testing is reduced to $F(\boldsymbol{z}) = \frac{\sum_{k=1}^{K}(z_k - \bar{z})^2}{(K-1)A}$, and large $F(\boldsymbol{z})$ tends to reject H_0 [13]. We can rewrite the JS estimator in a more insightful way,

$$\rho^{JS} = \min\left(1, \frac{K-3}{(K-1)F(\boldsymbol{z})}\right). \tag{17}$$

If H_0 is true, $F(\boldsymbol{z})$ should be small which leads to large ρ^{JS}, then JS heavily shrinks z_k toward $\bar{z}$. Note that the best estimator of θ_k when H_0 is true is exactly the grand mean $\bar{z}$. Hence, JS estimator implicitly conduct the hypothesis testing (16) to decide the shrinkage factor, $1 - \rho^{JS}$. With the aid of JS estimator, the parametric and nonparametric tail probability estimator can be redesigned as in Fig. 4 and Fig. 5.

VI. General Shrinkage Estimators

To obtain better estimators beyond JS, we let $Z_k \overset{\text{indep.}}{\sim} N(\theta_k, A_k)$ such that Z_k's may have unequal variances. Then we are interested in finding the optimal method in a general class of shrinkage estimators.

Definition 2. *General Shrinkage Estimators*
Let $\mathcal{S}$ be the set of all shrinkage estimators satisfy, $\forall k$, $\hat{\theta}_k = (1 - b_k)z_k + b_k c$, where $b_k \in [0,1]$ and $|c| \le \max_k |z_k|$.

Note that both conventional maximum likelihood method and JS estimators are in $\mathcal{S}$. Moreover, if we apply the conjugate hierarchical model,

$$Z_k|\boldsymbol{\theta} \overset{indep.}{\sim} N(\theta_k, A_k), \quad \Theta_k \overset{i.i.d.}{\sim} N(\tau, \eta). \tag{18}$$

The Bayesian MMSE estimator knowing τ and η,

$$\hat{\theta}_k^B = (1 - \rho_k^B)z_k + \rho_k^B \tau, \quad \rho^B = \frac{A_k}{\eta + A_k}, \tag{19}$$

is also in $\mathcal{S}$. Estimations of τ and η are used in (19) in empirical Bayesian due to the unknown of τ and η. Again, estimators derived in empirical Bayesian are in $\mathcal{S}$. Therefore, $\mathcal{S}$ is a general class of shrinkage estimators. The next problem is to determine $(\boldsymbol{b}, c)$ such that the derived estimator has good risk property.

A. Asymptotic Optimal Shrinkage Estimator

Definition 3. *Stein's Unbiased Risk Estimator (SURE)*
An unbiased risk estimator of $\mathcal{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ when $\hat{\boldsymbol{\theta}} \in \mathcal{S}$ is

$$\text{SURE}(\boldsymbol{b}, c) = \frac{1}{K}\sum_{k=1}^{K}[b_k^2(z_k - c)^2 + (1 - 2b_k)A_k]. \tag{20}$$

Definition 4. *SURE Shrinkage Estimator [11]*
Define the monotonicity constraint (MON): $b_i \le b_j$ if $A_i \le A_j$. The SURE shrinkage estimator is $\hat{\boldsymbol{\theta}}^s \in \mathcal{S}$ with $(\boldsymbol{b}, c)$ determined by the optimization,

$$\begin{array}{ll} \text{minimize} & \text{SURE}(\boldsymbol{b}, c) \\ \text{subject to} & \boldsymbol{b} \in [0,1]^K,\ |c| \le \max_k |z_k|,\ \textit{MON}. \end{array} \tag{21}$$

Theorem 2. *Asymptotic Optimality [11]*
If the regularity conditions

1) $\limsup_{k\to\infty} \sum_{i=1}^{k} A_i^2 < \infty,\ \limsup_{k\to\infty} \sum_{i=1}^{k} A_i\theta_i^2 < \infty.$
2) $\exists \delta > 0,\ \limsup_{k\to\infty} \sum_{i=1}^{k} |\theta_i|^{2+\delta} < \infty.$

hold, we have

$$\limsup_{K\to\infty}[\mathcal{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^S) - \mathcal{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})] \le 0, \forall \hat{\boldsymbol{\theta}} \in \mathcal{S}. \tag{22}$$

Intuitively, we want to find the estimator with minimum risk. However, since $\boldsymbol{\theta}$ is unknown, so is the true risk $\mathcal{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$. We cannot derive an estimator by directly minimize the true risk. Fortunately, (20) is a good estimation of the true risk as $K \to \infty$, hence minimizing SURE leads to estimators with small risk. Furthermore, (22) shows that $\hat{\boldsymbol{\theta}}^s$ is actually asymptotic optimal in the class $\mathcal{S}$. The identifier using SURE shrinkage estimator is shown in Fig. 6.

B. Water-Filling Algorithm

To implement SURE shrinkage estimator we proposed a water-filling type algorithm to solve the multi-dimensional optimization (21). For a fixed c and let $\boldsymbol{b}(c)$ be the minimizer of (21), we can transform (21) to the 1-dimensional optimization

$$\underset{|c| \le \max_k |z_k|}{\text{minimize}} \quad \text{SURE}(\boldsymbol{b}(c), c). \tag{23}$$

Therefore, it suggests us to propose an algorithm inspired by water-filling to obtain $\boldsymbol{b}(c)$. Fixed c and neglect the MON at the beginning. Let $d_k = (z_k - c)^2$ and $b_k^*(c) = \frac{A_k}{(z_k - c)^2}$, then

$$\boldsymbol{b}(c) = \arg\min_{\boldsymbol{b}\in[0,1]^K} \sum_{k=1}^{K} d_k(b_k - b_k^*(c))^2 \tag{24}$$

$$= \min(1, b_k^*(c)). \tag{25}$$

However, we need to adjust $b_k(c)$ such that the MON is satisfied. Without loss of generality, let $A_i \le A_j$ if $i < j$ then the MON is simplified to $b_i(c) \le b_j(c)$ if $i < j$. Note that if $b_k(c) \ge b_{k+1}(c)$, $b_{k+1}(c)$ and $b_k(c)$ must be adjusted to the same value to attain the MON. Therefore, we group

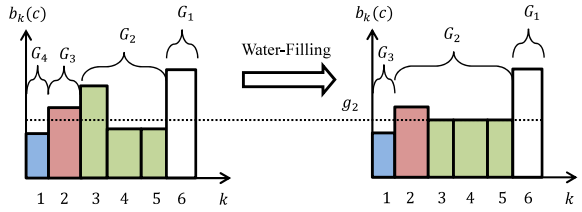

Fig. 7. Illustration of **Algorithm 1.** At the beginning, there are four groups in left plot. After one adjustment, $b_k(c)$'s in group G_2 have the same value g_2 as derived in (27), and there remains three groups in the right plot.

Algorithm 1 Water-Filling Type Algorithm

1: $b_k(c) \leftarrow \min(1, b_k^*(c)),\ \forall k = 1, 2, \ldots, K.$
2: **while** MON is not satisfied **do**
3: Grouping with $\mathcal{G} = \{G_i\}$ as in **Definition 5**.
4: **for** $G_i \in \mathcal{G}$ **do**
5: $b_k(c) \leftarrow \frac{\sum_{j\in G_i} d_j b_j^*}{\sum_{j\in G_i} d_j}, \forall k \in G_i.$
6: **end for**
7: **end while**
8: **return** $b_k(c),\ \forall k = 1, 2, \ldots, K.$

those equal values $b_k(c)$'s together and adjust their values to achieved the MON.

Definition 5. *Grouping*
A grouping $\mathcal{G} = \{G_i\}$ is a partition of $\{1, 2, \ldots, K\}$ such that $k, k+1 \in G_i$ if $b_k(c) \geq b_{k+1}(c)$ for all $k = 1, 2, \ldots, K-1$.

let g_i be the adjusted value of all $b_k(c)$ when $k \in G_i$, and $r_k(c) = A_k + \frac{A_k^2}{(z_k - c)^2}$. The value contributed by G_i to (20) is

$$\text{SURE}_i(g_i, c) = \frac{1}{K} \sum_{k\in G_i} \left[d_k (g_i - b_k^*(c))^2 + r_k(c) \right]. \quad (26)$$

Minimize (26) with respect to $g_i \in [0, 1]$, we get

$$g_i = \min\left(1, \frac{\sum_{k\in G_i} d_k b_k^*(c)}{\sum_{k\in G_i} d_k}\right). \quad (27)$$

After iterations as shown in Fig. 7, **Algorithm 1** inspired by water-filling leads to $b_k(c)$'s satisfying the MON. Now, we can implement SURE shrinkage estimator by solving (23) instead of (21).

VII. Simulations

Since all identifiers share the same preprocessor, we only compare tail probability estimators in Fig. 2-6. Major comparisons of interest are 1) asymptotic performance in K, 2) Robustness against t_{out} and 3) Robustness against signal model.

A. Simulation Settings

We use $\sigma_k^2 \overset{i.i.d.}{\sim} U(0.5, 1)$ in all simulations. In Fig. 8,9,11,12, we use $\mu_k^2 \overset{i.i.d.}{\sim} U(0.5, 1)$ such that all sources have diverse signal types. In Fig. 10 and 13, we let $\mu_k = 2 - \sigma_k^2/2$ such that all source signals have the same mean.

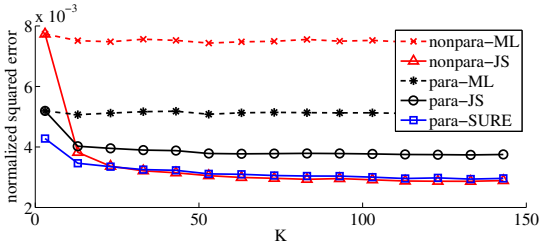

Fig. 8. Asymptotic behaviors of tail probability estimators with small samples, $N = 30$. All shrinkage based methods are benefit from the increase of K with fast convergence rate. The improvement of nonpara-JS is dramatic especially in this small samples case.

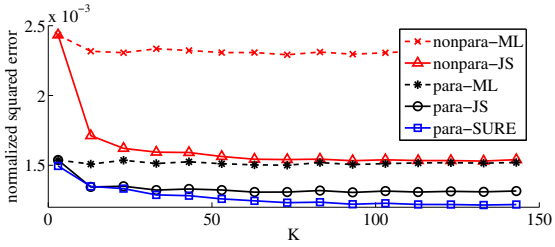

Fig. 9. Compare with Fig. 8, for large sample size $N = 100$, all parametric methods especially those using shrinkage estimators outperform nonparametric methods.

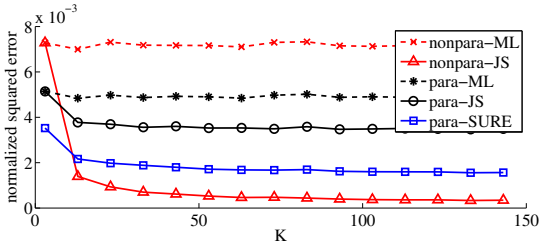

Fig. 10. For samples of small size $N = 30$ and sources with equal mean signals. Compare with Fig. 8, performance of nonpara-JS and para-SURE become better, especially the nonpara-JS.

Hence, we can compare methods under different mechanisms of signal generation. We fixed $t_{\text{out}} = \exp(1)$ in Fig. 8 to 10 to examine the asymptotic performance with different sample sizes. In Fig. 11 to 13, we fixed $K = 50$ to compare the performance over t_{out} with different sample sizes.

B. Simulation Results

Fig. 8-10 show that shrinkage based systems benefit from the increase of K. And the fast convergence rate makes shrinkage systems effective even when K is not actually large. Among parametric systems, the para-SURE eventually outperforms the others as K increase, which is guaranteed by its asymptotic optimality. The improvement from nonpara-ML to nonpara-JS are tremendous, which makes nonpara-JS even better than parametric methods in small sample cases. The reason is that only $\boldsymbol{\pi}$ to be estimated in nonparametric system, while there are $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ in parametric methods. Hence, parametric methods should pay the price of inaccurate estimation of σ_k^2 especially when N is small. As N increases,

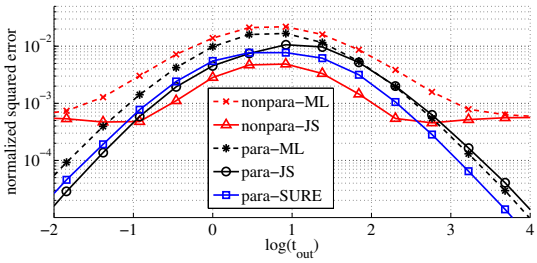

Fig. 11. With extreme small sample size $N = 10$, the nonpara-JS is the best for a moderate range of t_{out}. And the superiority of shrinkage based systems are uniformly in t_{out}.

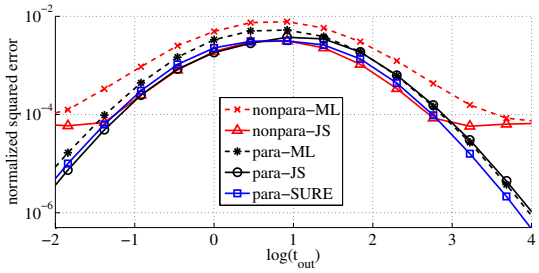

Fig. 12. Compare with Fig. 11, when $N = 30$, the nonpara-JS has similar performance with para-SURE for a wider range of t_{out}.

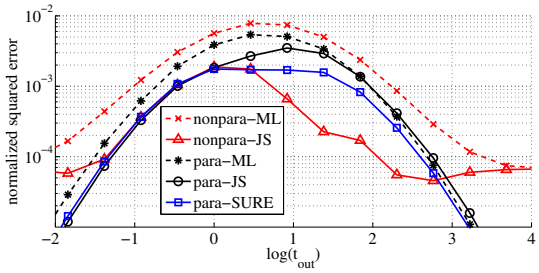

Fig. 13. For $N = 30$ and sources with equal mean signals, nonpara-JS outperforms the others for a wide range with incomparable performance.

all parametric methods eventually get better than nonparametric ones as shown in Fig. 9.

Although performance depends on t_{out}, the uniform superiority of shrinkage systems over all t_{out} are shown in Fig. 11-13. Surprisingly, in certain range of t_{out}, nonpara-JS beats the others when the sample size is small. Note that nonparametric methods need to pay the price of unknown parametric model of signal when t_{out} is relative small or large. However, the working range of nonparametic methods are wide enough for most real world applications. Consider the possible strongest outlier with $\mu_k = 1$ and $\sigma_k^2 = 1$, it only exceeds $t_{\text{out}} = \exp(3)$ with small probability 0.0228. Hence, the working range of nonparametirc methods is wide enough for general situations.

The robustness of nonparametric methods against signal type can be revealed by comparing Fig. 9 and 12 to Fig. 10 and 13. When sources have equal mean signals, the superiority of nonpara-JS becomes even stronger. It is because that parametric methods do not rely on any assumption of signal distribution, while parametric methods do not well utilize the dependence between μ_k and σ_k^2.

In general, the fast convergence of performance in K is a substantial support of our approach in large-scale datasets, when K is large. For moderate range of t_{out}, the nonpara-JS is highly suggested due to its effectiveness and easy implementation especially when N is small. In the other hand, with enough N and performance has the highest priority, the para-SURE is the best choice.

VIII. Conclusion

In this paper, we systematically develop outlier identifiers using shrinkage estimators for sensor network data. Simulations show the effectiveness and efficiency of the proposed methods especially in large-scale limited-sample cases, which is frequently encountered in many applications like environmental monitoring. Moreover, we proposed a water-filling type algorithm to obtain the asymptotic optimal shrinkage estimators, and thus applicable to general applications of data analysis. At the end, We emphasize that our methods are developed under general assumptions. Therefore, this work can be the foundation to resolve a class of outlier detection for sensed data.

References

[1] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 12, no. 2, pp. 159–170, 2010.

[2] A. Tajer, V. V. Veeravalli, and H. V. Poor, "Outlying sequence detection in large data sets: A data-driven approach," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 44–56, Sept 2014.

[3] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250–2267, Sept 2014.

[4] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar 1986.

[5] X. Wang and H. V. Poor, "Blind multiuser detection: a subspace approach," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 677–690, Mar 1998.

[6] W.-C. Wu and K.-C. Chen, "Identification of active users in synchronous cdma multiuser detection," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 9, pp. 1723–1735, Dec 1998.

[7] W. James and C. Stein, "Estimation with quadratic loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 361–379, 1961.

[8] C. N. Morris, "Parametric empirical bayes inference: Theory and applications," *Journal of the American Statistical Association*, vol. 78, no. 381, pp. 47–55, 1983.

[9] B. Efron and C. Morris, "Data analysis using stein's estimator and its generalizations," *Journal of the American Statistical Association*, vol. 70, no. 350, pp. 311–319, 1975.

[10] G. Casella, "An introduction to empirical bayes data analysis," *The American Statistician*, vol. 39, no. 2, pp. 83–87, 1985.

[11] X. Xie, S. C. Kou, and L. D. Brown, "Sure estimates for a heteroscedastic hierarchical model," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1465–1479, 2012.

[12] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 387–392, Apr 1985.

[13] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Thomson Learning, 2002.

[14] F. J. Anscombe, "The transformation of poisson, binomial and negative-binomial data," *Biometrika*, vol. 35, no. 3-4, pp. 246–254, 1948.