Mean estimation for entangled single-sample distributions

Ankit Pensia*, Varun Jog†, and Po-Ling Loh*‡
Departments of Computer Science*, Electrical & Computer Engineering†, and Statistics‡
University of Wisconsin - Madison
Email: {pensia, vjog, polingloh}@wisc.edu

***Abstract*—We consider the problem of estimating the common mean of univariate data, when independent samples are drawn from non-identical symmetric, unimodal distributions. This captures the setting where all samples are Gaussian with different unknown variances. We propose an estimator that adapts to the level of heterogeneity in the data, achieving near-optimality in both the i.i.d. setting and some heterogeneous settings, where the fraction of "low-noise" points is as small as $\frac{\log n}{n}$. Our estimator is a hybrid of the modal interval, shorth, and median estimators from classical statistics. The rates depend on the percentile of the mixture distribution, making our estimators useful even for distributions with infinite variance.**

I. Introduction

Many modern data sets involve various forms of heterogeneity that lead to new challenges in estimation and prediction. Whereas much of classical statistics focuses on convergence guarantees for i.i.d. observations, both the independence and identical distribution assumptions may be called into question in specific scientific applications [1]–[5].

We focus on the problem of estimating a common mean when univariate data are assumed to be generated independently, but from non-identical distributions. The special case where each sample is drawn from a normal distribution with a potentially different variance was studied by Chierichetti et al. [6], who showed the existence of a gap between estimation error rates of the optimal estimator when both the set of variances and their assignments are known (given by the maximum likelihood estimator) and the best possible estimator in the case where the variances of the distributions are completely unknown. Furthermore, Chierichetti et al. [6] presented a mean estimator in the unknown variance case based on calculating the "shortest gap" between samples, and derived upper bounds on the estimation error of their algorithm that generally behave better than the naive mean estimator (which has estimation error rates depending on the maximum variance) or the median estimator (which is suboptimal when a very large number of observations are drawn from distributions with large variances).

As discussed in Chierichetti et al. [6], a practical motivation for analyzing this problem is aggregation of user ratings in crowdsourcing, where the rating reported by each user might be drawn from a distribution centered around the true quality of the item, but with a different variance corresponding to the expertise of the user. Importantly, only one observation is available from each distribution, although the aggregate data are drawn from a Gaussian mixture.

A natural question is whether the estimators proposed by Chierichetti et al. [6] might also perform well in non-Gaussian settings. For instance, one might ask whether concentration inequalities for sub-Gaussian random variables might be plugged into the analysis in natural ways to obtain good upper bounds. Furthermore, although Chierichetti et al. [6] derive lower bounds for the behavior of the best possible estimator in the unknown variance setting, their work leaves open the question of whether their proposed estimator actually performs optimally, and for which collections of variances.

In this paper, we revisit the problem of common mean estimation and substantially generalize the setting beyond Gaussian mixtures. In particular, the only assumption we impose on each of the component distributions is symmetry and unimodality about a common mean. Although our proposed estimator is similar to the estimator studied by Chierichetti et al. [6], we use a rather different approach for the analysis, which allows us to obtain bounds without assuming Gaussianity, sub-Gaussianity, or even finite variances of individual distributions. Our analysis is inspired by ideas in empirical process theory, and the upper bounds involve percentiles of the overall mixture distribution, making them useful even in the case of Cauchy-type distributions with heavy tails.

Our proposed estimators are connected to classical estimators appearing in the statistics literature, notably the modal interval estimator [7] and the shorth estimator [8]. However, existing analysis of these estimators has generally been asymptotic and limited to i.i.d. data. In fact, it is well-known that in the i.i.d. setting, both the modal interval and shorth estimators have an $n^{-\frac{1}{3}}$ convergence rate [9], compared to the faster $n^{-\frac{1}{2}}$ convergence rate of the sample mean—on the other hand, our analysis shows that these estimators enjoy superior performance to the sample mean when a substantial fraction of the component distributions have variances that are extremely large, or even infinite. This underscores the fundamental fact that estimators which are suboptimal in a "clean" data setting may be preferable from the point of view of robustness.

Parameter estimation of mixture models is well-studied in statistics and computer science [10]–[14]. However, our setting is somewhat different from the canonical setting, since the number of observations is not large relative to the number of component distributions; rather, the number of components in

the mixture could be as large as the number of observations. On the other hand, the parameters of the component mixtures are fundamentally entangled via a common mean parameter, which is the quantity we wish to estimate. Consequently, although much of the literature in statistical estimation for mixture models requires the underlying component distributions to possess certain tail characteristics such as Gaussianity or log-concavity, such assumptions are not necessary to obtain small estimation error in our setting.

The remainder of the paper is organized as follows: In Section II, we define notation and the basic estimators we will consider, which are analyzed individually in Section III. In Section IV, we combine the ideas from the previous sections to present a hybrid estimator and derive an upper bound on the final output of the estimation algorithm. We also discuss a setting in which the performance of our algorithm is nearly optimal. Proofs are omitted due to space considerations.

Notation: For two real-valued functions $f(n)$ and $g(n)$, we write $f(n) = \omega(g(n))$ if for every real constant $c > 0$, there exists $n_0 \geq 1$ such that $f(n) > c \cdot g(n)$ for every integer $n \geq n_0$. We write w.h.p., or "with high probability," to mean with probability tending to 1 as the sample size increases. We use C and c to represent absolute positive constants Similarly, we use C_t to represent a positive number that depends only on t.

II. Preliminaries

For $x \in \mathbb{R}$ and $r \geq 0$, let $f_{x,r}$ denote the indicator function of the interval $[x - r, x + r]$. Let

$$\mathcal{H} := \{f_{x,r} : x \in \mathbb{R}, r \in \mathbb{R}, r \geq 0\},$$
$$\mathcal{H}_r = \{f_{x,r'} : x \in \mathbb{R}, r' \in \mathbb{R}, 0 \leq r' \leq r\}.$$

Both $\mathcal{H}$ and $\mathcal{H}_r$ have VC dimension 2 [15].

Throughout the paper, we will assume that $X_i \sim P_i$ independently, where each P_i has a density. All the P_i's are assumed to be symmetric and unimodal around a common median μ^*. Furthermore, the P_i's are unknown to the learner and need not even be from the same parametric family of distributions. Let q_i and σ_i be the interquartile range and standard deviation of P_i, respectively. Recall that the interquartile range satisfies $\mathbb{P}(|X_i - \mu^*| \leq q_i) = \frac{1}{2}$. We use $q_{(i)}$ and $\sigma_{(i)}$ to denote the i^{th} smallest interquartile range and standard deviation, respectively. For simplicity of presentation, we will assume that $\mu^* = 0$. There is no loss of generality, since the estimators that we consider are translation invariant. Thus, the error of an estimator $\widehat{\mu}$ will be measured by $|\widehat{\mu}|$.

For a function f, we use $R_n(f) := \frac{1}{n}\sum_{i=1}^n f(X_i)$ to denote the expectation of f with respect to the empirical distribution of $X_1, \ldots, X_n$. Let

$$R(f) := \frac{1}{n}\sum_{i=1}^n \mathbb{E}\, f(X_i).$$

Thus, $R(f)$ is the expectation of f with respect to the mixture $\overline{P} := \frac{1}{n}\sum_{i=1}^n P_i$, which is again unimodal and symmetric.

A. Properties of $\overline{P}$

Since we are not given access to the individual P_i's, we will argue about the problem through the lens of $\overline{P}$. We first establish some useful properties. Let

$$R_r^* := \sup_{f \in \mathcal{H}_r} R(f) = R(f_{0,r}),$$

where the second equality follows by symmetry and unimodality. It is also equal to the probability of the interval $[-r, r]$ under $\overline{P}$.

Lemma 1: We have the following properties:

(i) For any $r > 0$ and $x, x' \in \mathbb{R}$, if $|x| < |x'|$, then $R(f_{x,r}) \geq R(f_{x',r})$.
(ii) For any $x \in \mathbb{R}$, if $r < r'$, then $R(f_{x,r}) \leq R(f_{x,r'})$.
(iii) If $0 < r < r'$, then $\frac{R_r^*}{r} > \frac{R_{r'}^*}{r'}$.
(iv) If $0 < r < r'$, then $R(f_{r',r}) < \frac{r}{r'} R_{r'}^*$.
(v) If $1 \leq k \leq n$, then $\frac{k}{n} < R_{q_{(2k)}}^*$ and $\frac{k}{n} < R_{2\sigma_{(2k)}}^*$.

The proofs proceed using simple calculus and algebraic manipulations, relying only on the properties of symmetry and unimodality. Lemma 1 shows that we can use $\overline{P}$ as a measure of distance between two intervals. In particular, if two intervals with the same center/radius are close under R, the respective radii/centers must also be close.

B. Estimators

We now define the estimators that will serve as building blocks for our algorithms. All of these estimators can be implemented efficiently after sorting the data points.

Estimator 1 (r-modal interval): The r-modal interval outputs the center of the most populated interval of length r:

$$\widehat{\mu}_{M,r} := \operatorname*{argmax}_x R_n(f_{x,r}).$$

Estimator 2 (k-shortest gap / shorth estimator): The k-shortest gap estimator, $\widehat{\mu}_{S,k}$, outputs the center of the shortest interval containing at least k points. More precisely, we define

$$\widehat{r}_k := \inf\left\{r : \sup_x R_n(f_{x,r}) \geq \frac{k}{n}\right\}, \qquad \widehat{\mu}_{S,k} := \widehat{\mu}_{M,\widehat{r}_k}.$$

The traditional shorth estimator [8], [9] corresponds to $k = 0.5n$.

We also define the population-level quantities

$$r_k := \inf\left\{r : \sup_x R(f_{x,r}) \geq \frac{k}{n}\right\} = \inf\left\{r : R(f_{0,r}) \geq \frac{k}{n}\right\},$$

where the last equality follows from unimodality and symmetry. Note that r_k measures the spread of $\overline{P}$ and $r_{0.5n}$ is the interquartile range of $\overline{P}$. Furthermore, since $\overline{P}$ has a density, we have $R_{r_k}^* = \frac{k}{n}$. In particular, by Lemma 1(v), we have $r_k \leq q_{(2k)}$ and $r_k \leq 2\sigma_{(2k)}$, although these bounds may be loose (for instance, r_k could be finite even if $\sigma_{(1)}$ is infinite). However, we are guaranteed that r_k will be small if $2k$ points come from "nice" distributions.

The k-median outputs the centermost k points of the data. Note that the output is a set rather than a point estimator; however, the k-median will be useful as a preprocessing step

before applying the modal interval or shorth estimators, to obtain better estimation error rates.

Estimator 3 (k-median): The k-median outputs a subset S_k that contains the centermost k points. More precisely, we have $X_i \in S_k$ if and only if $\widehat{\theta}_{\text{med},-k} \le X_i \le \widehat{\theta}_{\text{med},k}$, where

$$\widehat{\theta}_{\text{med},k} := \inf\left\{\theta : \psi_n(\theta) \ge \frac{k}{n}\right\},$$
$$\widehat{\theta}_{\text{med},-k} := \sup\left\{\theta : \psi_n(\theta) \le \frac{-k}{n}\right\},$$

and $\psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \text{sign}(\theta - X_i)$. The sample median corresponds to taking $k = 0$.

III. Performance analysis of individual estimators

We use the following concentration bound to prove tight concentration results for the modal and the shorth estimators.

Theorem 1: For any fixed $t \in (0, 1]$ and $n > 1$, we have

$$\mathbb{P}\left\{\sup_{f\in\mathcal{H}_r} |R_n(f) - R(f)| \ge tR_r^*\right\} \le 2\exp\left(-cnR_r^* t^2\right),$$

provided r is large enough so that $R_r^* \ge \frac{C_t \log n}{n}$, where $C_t = \left(\frac{144}{t}\right)^2$ and $c = \frac{1}{200}$.

The proof modifies the VC-type bounds derived for i.i.d. settings. This theorem is useful because the bounds are adaptive to the problem, compared to the traditional $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ distribution-independent bound. However, note that Theorem 1 requires the mass R_r^* lying around the true mode to be sufficiently large.

A. Modal interval

The following theorem provides a high-probability bound on the error of the modal interval estimator:

Theorem 2: Let r be such that $R_r^* \ge C_{0.25}\left(\frac{\log n}{n}\right)$. Then with probability at least $1 - 2\exp(-cC_{0.25}\log n/16)$, we have

$$|\widehat{\mu}_{M,r}| \le \frac{2r}{R_r^*}. \tag{1}$$

The proof of Theorem 2 proceeds by using Theorem 1 to bound the ratio between $R(f_{\widehat{\mu}_{M,r},r})$ and R_r^*, and then using Lemma 1 to turn this into a deviation bound on $|\widehat{\mu}_{M,r}|$.

Remark 1: The bound in Theorem 2 is tighter for smaller values of r by Lemma 1 (iii). Thus, the choice of r which optimizes the bound satisfies $R_r^* = C_{0.25}\left(\frac{\log n}{n}\right)$, yielding the bound $|\widehat{\mu}_{M,r}| \le \frac{2nr_{C_{0.25}\log n}}{C_{0.25}\log n}$. However, it poses a challenge because we do not know P: If r is too small, then R_r^* might not be large enough and the bounds might not hold, whereas if r is too large, then the resulting bound is loose.

Fortunately, an estimator with near-optimal performance may be obtained via Lepski's method [16]. The basic steps are as follows: Define $r^* = r_{C_{0.25}\log n}$ to be the interval width satisfying $R_{r^*}^* = C_{0.25}\left(\frac{\log n}{n}\right)$, and suppose we have rough initial estimates $r_{\min}$ and $r_{\max}$ such that $r_{\min} \le r^* \le r_{\max}$. Define $r_j := r_{\min}2^j$, and define

$$\mathcal{J} := \{j \ge 1 : r_{\min} \le r_j < 2r_{\max}\}.$$

We then define the index j_* to be

$$\min\left\{j \in \mathcal{J} : \forall i > j \text{ s.t. } i \in \mathcal{J}, |\widehat{\mu}_{M,r_i} - \widehat{\mu}_{M,r_j}| \le \frac{4r_i}{R_{r_i}^*}\right\},$$

which may be calculated using pairwise comparisons of the modal interval estimator computed over the gridding of $[r_{\min}, r_{\max}]$. We define $j_* = \infty$ if the set is empty; as proved in the theorem below, we have $j_* < \infty$, w.h.p. We then have the following result:

Theorem 3: With probability at least $1 - 2\left(1 + \log_2\left(\frac{2r_{\max}}{r_{\min}}\right)\right)\exp(-Cnk/8)$, we have

$$|\widehat{\mu}_{M,r_{j_*}}| \le \frac{12nr^*}{C_{0.25}\log n}. \tag{2}$$

Note that the cost of using Lepski's method is a factor of 6 in the estimation error. Of course, the validity of the method requires the availability of the rough bounds $r_{\min}$ and $r_{\max}$. A natural way to obtain rough bounds on r^* from the data is to use the shortest gap estimator, which returns the shortest interval containing at least $C_{0.25}\log n$ points. We can again use Theorem 1 and multiplicative form of the Chernoff bound [17] to show that $\widehat{r}_k$ does not fluctuate too wildly from its empirical counterpart:

Lemma 2: For $k \ge 4C_{0.5}\log n$, with probability at least $1 - \exp(-k/8) - \exp(-ck/16)$, we have $r_{k/2} \le \widehat{r}_k \le r_{2k}$.

Accordingly, we may use

$$r_{\min} = \hat{r}_{C_{0.25}\log n/2},$$
$$r_{\max} = \hat{r}_{2C_{0.25}\log n}.$$

Finally, note that the modal interval estimator $\widehat{\mu}_{M,r}$ is an M-estimator [18]:

$$\widehat{\mu} \in \arg\min_{\mu}\left\{\frac{1}{n}\sum_{i=1}^n g(x_i - \mu)\right\},$$

with loss function $g(x) = -f_{0,r}(x)$. Clearly, the loss is non-convex; however, the modal interval estimator is nonetheless computable, since the overall objective function is piecewise constant, with transitions lying only at the n data points. Whereas a more straightforward analysis of convex M-estimators (e.g., using a Huber loss) would yield somewhat similar bounds on estimation error, such arguments would generally require tail assumptions on the component distributions $\{P_i\}$, which we avoid altogether in our analysis of the modal interval estimator.

B. Shorth

Guarantees for the shorth estimator are similar to the modal interval estimator, but computing the shorth does not require an extra step for adapting to the width of the optimal interval. We have the following theorem, proved using Theorem 1 and Lemma 2:

Theorem 4: Suppose $k \geq \max\{4C_{0.5}, C_{0.25}\}\log n$. With probability at least $1 - \exp(-ck/16) - \exp(-k/8)$, we have

$$|\widehat{\mu}_{S,k}| \leq \frac{4nr_{2k}}{k} < \frac{4n\min\left(q_{(2k)}, 2\sigma_{(2k)}\right)}{k}.$$

Remark 2: Lemma 1(iii) shows that the upper bound is actually tighter for small k: for $k' > k$, we have $kr_{2k'} > k'r_{2k}$. The smallest value permissible from our theory would be $k = \Omega(\log n)$. Also note that the upper bound in Theorem 4 for the shorth estimator resembles the bound in Theorem 3, except for the fact that the bound for the modal interval estimator involves the quantity $r_{C_{0.25}\log n}$ rather than $r_{2C_{0.25}\log n}$, and the latter could be somewhat larger depending on the spread of $\overline{P}$. Furthermore, both upper bounds in Theorem 4 may sometimes be loose: In particular, if the X_i's were i.i.d. then r_{2k} would be of order $\Theta\left(\frac{k}{n}\right)$, so the bound $\frac{nr_{2k}}{k}$ would be of constant order, whereas it is known [9] that the shorth estimator is consistent for $k = 0.5n$.

We also discuss an example describing a phase transition in the performance of the shorth estimator with respect to variances. The proof of Proposition 1 uses Theorem 4 and Theorem 1, along with properties of Gaussian distributions.

Example 1: Let $\alpha > 0$ be a constant. Let $X_i \sim \mathcal{N}(0, \sigma_i^2)$, independently, where $\sigma_i = 1$ for $i \leq C\log n$ and $\sigma_i = n^\alpha$ otherwise, for some large constant C.

Proposition 1: For Example 1 and $k = 0.5C\log n$, w.h.p., we have $|\widehat{\mu}_{S,k}| = \mathcal{O}(n^\alpha)$ for $\alpha < 1$ and $|\widehat{\mu}_{S,k}| = \mathcal{O}(1)$ for $\alpha \geq 1$.

Remark 3: Let $\overline{P}_n$ be the empirical distribution of $X_1, \ldots, X_n$. Both the shorth and the modal estimator are "local" estimators that only consider the value of $\overline{P}_n$ in small windows. As we increase the variance of noisy points, the distribution $\overline{P}$ approaches 0 around μ^*. The shorth and modal interval estimators make mistakes when $\overline{P}$ is flat after normalization, meaning that the density at $x + \mu^*$ is within a $(1-\epsilon)$-factor of its density at μ^*, for $\epsilon = o(1)$. If this is the case, then $\overline{P}_n$ might assign higher mass at $x + \mu^*$ than μ^* due to stochasticity introduced by sampling, so a local method would mistakenly choose $x + \mu^*$ over μ^*. If an adversary tried to alter the estimator by making the variance of the points very high ($\alpha \gg 1$), then although $\overline{P}$ would approach 0, the normalized density would not be flat. An extreme example of this can be seen when variance of noisy points is "∞": Near μ^*, the distribution $\overline{P}$ would behave like $\mathcal{N}(\mu^*, 1)$ scaled by $\mathcal{O}\left(\frac{\log n}{n}\right)$, which is *not* flat after normalization, although $\overline{P}$ approaches 0 very rapidly, so that the mean or median would behave poorly. As Proposition 1 shows, the shorth estimator would only suffer $\mathcal{O}(1)$ error in this case.

Finally, note that in Example 1, the sample median and even the mean would have an error of $\mathcal{O}(n^{\alpha-0.5})$. When $\alpha < 1$, this is noticeably better than the shorth estimator, which motivates the hybrid estimator proposed in Section IV.

C. k-median

For the median estimator, we have the following result:

Lemma 3: If $R^*_\epsilon \geq \frac{k}{n} + \delta$, then

$$\widehat{\theta}_{\text{med},k} \leq \epsilon \quad \text{and} \quad \widehat{\theta}_{\text{med},-k} \geq -\epsilon,$$

with probability at least $1 - 2\exp(-n\delta^2)$.

The proof proceeds by applying Hoeffding's inequality [15] on $\psi_n(\cdot)$ and noting that for $\theta > 0$, we have $\mathbb{E}\,\psi_n(\theta) = R^*_\theta$. We will be interested in the case when $k = \mathcal{O}(\sqrt{n})$ and $\delta = \mathcal{O}\left(\frac{\log n}{\sqrt{n}}\right)$. Lemma 3 states that if $f_{0,\epsilon}$ contains enough mass, the output of the k-median is bounded by ϵ. Compared to the shorth and modal interval estimators, the k-median is a more "global" estimator, because it looks at a much bigger interval $f_{0,\epsilon}$: When the latter $n - \mathcal{O}(\sqrt{n})$ points have very large variances, the density of $\overline{P}$ is much flatter, so the value of ϵ needed to contain $\frac{\log n}{\sqrt{n}}$ mass is quite large. On the other hand, recall that this is not an issue for the modal interval and shorth estimators, which only have error governed by the smallest $\mathcal{O}(\sqrt{n})$ variances.

IV. Hybrid estimator

We now present an algorithm that combines the shorth and k-median estimators in order to obtain superior performance for both fast and slow decay of $\overline{P}$. The algorithm computes the k_1-shorth estimator and k_2-median. If the shorth estimator lies within the median interval, the algorithm outputs the shorth; otherwise, it outputs the closest endpoint of the median interval. This estimator is similar to the estimator proposed by Chierichetti et al. [6] since it uses the median as a screening step. However, the shorth estimator is computed separately, bypassing the need for a delicate conditioning argument in the analysis.

Algorithm 1 Hybrid mean estimator

```
function HybridMeanEstimator(X_{1:n}, k_1, k_2)
    S_{k_1} ← kMedian(X_{1:n}, k_1).
    μ̂_{S,k_2} ← Shorth(X_{1:n}, k_2).
    if μ̂_{S,k_2} ∈ [min(S_{k_1}), max(S_{k_1})] then
        μ̂_{k_1,k_2} ← μ̂_{S,k_2}
    else
        μ̂_{k_1,k_2} ← closestPoint(S_{k_1}, μ̂_{S,k_2})
    end if
    return μ̂_{k_1,k_2}
end function
```

The following theorem provides an error bound for the hybrid estimator:

Theorem 5: If $k_1 = \sqrt{n}\log n$ and $k_2 = C_{0.25}\log n$, the output of the hybrid estimator in Algorithm 1 is bounded by

$$|\widehat{\mu}_{k_1,k_2}| \leq \frac{8\sqrt{n}\log n}{k_2} r_{2k_2} \leq \frac{8\sqrt{n}\log n}{k_2} \min\left(q_{(2k_2)}, 2\sigma_{(2k_2)}\right),$$

with probability at least $1 - 2\exp(-c'k_2) - 2\exp(-\log^2 n)$. The proof of Theorem 5 proceeds by considering two cases for $R^*_{\frac{8\sqrt{n}\log n}{k_2} r_{2k_2}}$, and uses Lemma 3 and Theorem 4. Importantly, the bound in Theorem 5 is finite even for heavy-tailed distributions with infinite variance. Finally, note that in

Algorithm 1, we could replace the shorth estimator by the modal interval estimator with adaptively chosen interval width (cf. Section III-A) to obtain similar error guarantees.

Remark 4: The bounds achieved by our estimators are problem-dependent, since for a fixed k_2, the quantity r_{k_2} decreases as the fraction of clean points increases. Under mild regularity conditions, if k samples have variance 1, then $r_{k_2} = \mathcal{O}\left(\frac{k_2}{k}\right)$, leading to the bound $\mathcal{O}\left(\frac{\sqrt{n}\log n}{k}\right)$. Thus, the estimation error vanishes when $k = \omega(\sqrt{n}\log n)$. In particular, for n i.i.d. distributions, we obtain an $\mathcal{O}\left(\frac{\log n}{\sqrt{n}}\right)$ rate, which is within a $\log$ factor of the optimal rate $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ of the sample mean. On the other hand, the mean is clearly suboptimal when the variances vary widely. The benefit of our hybrid estimator is that it can achieve reasonable error guarantees in both the cases of i.i.d. and highly heterogeneous observations, without knowing the collection of variances a priori.

As the following example illustrates, we can derive optimality of the hybrid estimator in certain settings:

Example 2 (Example 1 in Chierichetti et al. [6]): Choose the number of samples $n > 0$, and choose $p \in \left(0, \frac{1}{\sqrt{2n}}\right)$. For each i, let σ_i be chosen i.i.d. according to the following distribution: with probability p, let it be equal to $p^2 n$; and with probability $1-p$, let it be equal to $1-p$. Conditioned on σ_i, let $X_i \sim \mathcal{N}(\mu^*, \sigma_i)$ be drawn independently.

By Lemma 4.4 in Chierichetti et al. [6], the optimal estimator when the σ_i's are known is

$$\mathbb{E}\,|\widehat{\mu} - \mu^*| = \Theta\left(\min(p^{3/2}n^{1/2}, n^{-1/2})\right),$$

provided $pn = \Omega(\log n)$. However, if the σ_i's are unknown, Lemma 4.5 in Chierichetti et al. [6] states that if $pn = \Omega(\log n)$ and $p = o(n^{-1/2})$, any algorithm suffers $\mathbb{E}\,|\widehat{\mu} - \mu^*| = \Omega\left(\frac{1}{\sqrt{n}}\right)$. Note the gap in the optimal error between the two settings when $p = O(n^{-2/3})$.

However, Chierichetti et al. [6] do not establish the optimality of their proposed estimator. The following result, proved using Theorem 5, shows that our hybrid algorithm is indeed nearly optimal in this scenario, up to a logarithmic factor:

Proposition 2: If $p = \Omega\left(\frac{\log n}{n}\right)$ and $p = o(n^{-1/2})$ in Example 2, then the hybrid algorithm with $k_1 = \sqrt{n}\log n$ and $k_2 = \Theta(\log n)$ achieves $\mathbb{E}\,|\widehat{\mu}_{k_1,k_2} - \mu^*| = O\left(\frac{\log n}{\sqrt{n}}\right)$.

Note that the situation in Example 2 is in some sense a mild case of heterogeneity: It demonstrates a scenario where a provable gap exists between the convergence rates of estimators based on a priori knowledge of the σ_i's, but at the same time, the mean estimator would still achieve the optimal $O\left(\frac{1}{\sqrt{n}}\right)$ rate for estimators that do not rely on knowledge of the σ_i's. We leave the question of optimality of the hybrid estimator for a broader range of settings, dependent on the values of the σ_i's, for future work.

V. Conclusion

We have studied the problem of mean estimation of a heterogeneous mixture when the fraction of clean points tends to 0. We have shown that the modal interval and shorth estimator, which perform suboptimally in i.i.d. settings, are superior to the sample mean in such settings. We have also shown that these estimators and the k-median have complementary strengths that may be combined into a single hybrid estimator, which adapts to the given problem and is nearly optimal in certain settings. An important question for further study is whether the proposed hybrid estimator is always near-optimal, or optimal, for more general collections of variances.

Acknowledgment

AP and PL were partially supported by NSF grant DMS-1749857. PL thanks Gabor Lugosi for introducing her to the entangled mean estimation problem at the 2017 probability and combinatorics workshop in Barbados.

References

[1] R. Y. Liu, "Bootstrap procedures under some non-iid models," *The Annals of Statistics*, vol. 16, no. 4, pp. 1696–1708, 1988.

[2] M. Dundar, B. Krishnapuram, J. Bi, and R. B. Rao, "Learning classifiers when the training data is not iid." in *IJCAI*, 2007, pp. 756–761.

[3] I. Steinwart and A. Christmann, "Fast learning from non-iid observations," in *Advances in NIPS*, 2009, pp. 1768–1776.

[4] T. Zhu, P. Xiong, G. Li, and W. Zhou, "Correlated differential privacy: Hiding information in non-iid data set," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 2, pp. 229–242, 2015.

[5] S. R. Flaxman, D. B. Neill, and A. J. Smola, "Gaussian processes for independence tests with non-iid data in causal inference," *ACM Transactions on TIST*, vol. 7, no. 2, p. 22, 2016.

[6] F. Chierichetti, A. Dasgupta, R. Kumar, and S. Lattanzi, "Learning entangled single-sample Gaussians," in *Proceedings of the 25th Annual Symposium on Discrete Algorithms, SODA*, 2014, pp. 511–522.

[7] H. Chernoff, "Estimation of the mode," *Annals of the Institute of Statistical Mathematics*, vol. 16, no. 1, pp. 31–41, dec 1964.

[8] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey, *Robust Estimates of Location: Survey and Advances*. Princeton University Press, 1972.

[9] J. Kim and D. Pollard, "Cube root asymptotics," *Ann. Statist.*, vol. 18, no. 1, pp. 191–219, 03 1990.

[10] B. G. Lindsay, "Mixture models: Theory, geometry and applications," in *NSF-CBMS Regional Conference Series in Probability and Statistics*. JSTOR, 1995, pp. i–163.

[11] S. Dasgupta, "Learning mixtures of Gaussians," in *40th Annual Symposium on Foundations of Computer Science*. IEEE, 1999, pp. 634–644.

[12] S. Arora and R. Kannan, "Learning mixtures of arbitrary Gaussians," in *Proceedings of the 33rd annual ACM Symposium on Theory of Computing*, 2001, pp. 247–257.

[13] R. Kannan, H. Salmasian, and S. Vempala, "The spectral method for general mixture models," in *International Conference on Computational Learning Theory*. Springer, 2005, pp. 444–457.

[14] D. Achlioptas and F. McSherry, "On spectral learning of mixtures of distributions," in *International Conference on Computational Learning Theory*. Springer, 2005, pp. 458–469.

[15] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[16] O. V. Lepskii, "On a problem of adaptive estimation in Gaussian white noise," *Theory of Probability & Its Applications*, vol. 35, no. 3, pp. 454–466, 1991.

[17] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, 1st ed. Oxford University Press, 4 2016.

[18] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.