Proceedings of the 3rd International Conference on Statistics: Theory and Applications (ICSTA'21)
~~Prague, Czech Republic~~ Virtual Conference – August, 2021
Paper No. 135
DOI: 10.11159/icsta21.135

Robustness of Gaussian Mixture Reduction for Split-and-Conquer Learning of Finite Gaussian Mixtures

Qiong Zhang, Jiahua Chen
Department of Statistics, University of British Columbia
3182 Earth Sciences Building, 2207 Main Mall, Vancouver BC, Canada V6T 1Z4
qiong.zhang@stat.ubc.ca; jhchen@stat.ubc.ca

Abstract **-** In the era of big data, there is an increasing demand for split-and-conquer learning of finite mixture models. Recent work [1] proposes several split-and-conquer approaches for learning finite Gaussian mixtures and they are found to be both statistically and computationally efficient when the order of the mixture is correctly specified. Due to the nature of mixture models, correctly specifying the order of mixture on local machines can be an unrealistic assumption. In this paper, we evaluate the performance of several split-and-conquer learning approaches, both when the order is correct and when it is over-specified on the local machines, based on simulations. We find that there is a trade-off between robustness and computational efficiency: the computationally intensive approach is robust against over-specification, while the two computationally friendly approaches have compromised statistical performance when the order is over-specified. The results suggest that the information in the data about the true distribution is not lost in the split step of the learning, and aggregation strategies must be developed in a computationally and statistically efficient way.

Keywords**:** Finite Gaussian mixture, local machine, mixture reduction, split-and-conquer, transportation divergence.

1. Introduction

The split-and-conquer approach is an effective way of learning statistical models when the dataset is large or when the dataset is composed of many subsets that are stored on different local machines. A split-and-conquer approach consists of two steps: (i) local inference: standard inference is carried out on local machines; (ii) aggregation: the local results are transmitted to a central machine to be aggregated. Without relying on data sharing, such approaches have a built-in advantage in privacy consideration. The split-and-conquer approach has been successfully applied to learn generalized linear models [1], the kernel ridge regression model [2], and the local average regression models [3].

[1] investigates this approach under the finite Gaussian mixture model. Currently, they focus on the situation where the order of the mixture model is known and correctly specified. Every local machine learns a finite Gaussian mixture with the same and correct order. While the machine learning community has devoted most energy to this special case, it is of interest and great importance to develop the split-and-conquer approaches when the order of the mixture is potentially over-specified.

In this paper, we empirically evaluate the performance of a number of split-and-conquer approaches under this case. Our study reveals that there is a trade-off between robustness and computationally efficiency: the computationally intensive approach is robust against over-specification, while the two computationally friendly approaches have compromised statistical performance when the order is over-specified. We believe that the information in the data about the true model is not lost in the split step of the learning. Hence, there is a good promise to develop computationally friendly aggregation strategies to aggregate local estimates in a statistically efficient way. Full exploration of such remedies is left as future work.

The rest of this paper is organized as follows: we briefly introduce the split-and-conquer learning of Gaussian mixtures in Section 2. In Section 3, several split-and-conquer approaches are introduced. We study their performance under two finite Gaussian mixtures of order 3 in 1 and 2 dimensional spaces respectively. The preliminary conclusions based on the empirical study is given in Section 4.

2. Distributed Learning of Gaussian Mixtures

Let $\mathcal{F} = \{f(\cdot\,|\theta)\colon\ \theta \in \Theta\}$ be a parametric family of density functions with respect to some σ-finite measure, and $\mathrm{G} = \sum_{\mathrm{k}=1}^{\mathrm{K}} w_{\mathrm{k}}\{\theta_{\mathrm{k}}\}$ be a discrete probability measure on Θ that assigns probability w_k to subpopulation parameter θ_k, for some integer $K > 0$. The density function of a finite mixture on $\mathcal{F}$ is defined to be $f(x|\mathrm{G}) = \int f(x|\theta)\mathrm{dG}(\theta) = \sum_{\mathrm{k}=1}^{\mathrm{K}} w_{\mathrm{k}} f(x|\theta_{\mathrm{k}})$.

In this context, the measure G is called the mixing distribution, and the entries of $\mathbf{w} = (w_1, w_2, \cdots, w_K)^\tau$ are called the mixing weights. We denote the space of mixing distributions with order up to K by

$$\mathbb{G}_K = \left\{ \mathrm{G} = \sum_{k=1}^{K} w_k \{\theta_k\} | \theta_k \in \Theta, w_k \geq 0, \sum_{k=1}^{K} w_k = 1 \right\}.$$

A mixture of (exact) order K has its mixing distribution in $\mathbb{G}_K - \mathbb{G}_{K-1}$. When $\mathcal{F}$ is the family of Gaussian distributions such that

$$f(\boldsymbol{x}|\theta) = \phi(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^\tau \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\},$$

$\{f(x|\mathrm{G}) | G \sim \mathbb{G}_K\}$ is a finite Gaussian mixture model of order K. In this model, the subpopulation parameter space becomes $\Theta = \mathbb{R}^d \times \mathfrak{S}^d$ where $\mathfrak{S}^d$ is the space of symmetric and nonnegative definite matrices of dimension d. We use $\phi(\boldsymbol{x}|\mathrm{G})$ instead of $f(\boldsymbol{x}|\mathrm{G})$ for its density function.

Let $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ be a set of independent and identically distributed (IID) observations from a Gaussian mixture $\phi(\boldsymbol{x}|G)$ of order K. Suppose that $\mathcal{X}$ is partitioned into M subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_M$ completely at random and they are stored on M different local machines. Denote by N_m the sample size and $\lambda_m = N_m/N$ the proportion of samples stored on the mth machine. The local inference is performed by employing some learning strategy on the local machines and $\hat{G}_m$ of order K is an estimate of G based on subset $\mathcal{X}_m$, for $\mathrm{m} = 1, \ldots, \mathrm{M}$. The task of aggregation is to form an aggregated estimate of G through local estimates. One straightforward aggregated estimate is $\overline{\mathrm{G}} = \sum_{\mathrm{m}=1}^{\mathrm{M}} \lambda_m \hat{G}_m$. This is arguably a good estimate of the true mixing distribution in general. However, its induced mixture has inflated order even when the orders of all local estimates are correctly specified.

In view of the shortcoming of $\bar{G}$, we need a more deliberate aggregation step. One strategy is to reduce the order of $\phi(\boldsymbol{x}|\bar{G})$ from MK to K with minimum distortion. The problem of approximating a high order Gaussian mixture by one with a lower order is called Gaussian Mixture Reduction (GMR), see [5]–[9]. In particular, [1] proposes a method specifically targeting the GMR problem in learning Gaussian finite mixtures.

Let $D(\cdot \mid \cdot)$ be a divergence between two mixtures. Given a choice of $D(\cdot \mid \cdot)$, it is natural to search for a mixture with pre-designated order through

$$\bar{G}^R = \mathrm{argmin}_{\mathrm{G} \in \mathbb{G}_\mathrm{K}} \mathrm{D}\big(\phi(x|\overline{\mathrm{G}}) | \phi(x|\mathrm{G})\big). \tag{1}$$

This definition provides a principle but lacks specifications for the divergence. Divergences that facilitate both statistical and computational efficiency should be considered. The following definition of [1] is helpful toward this goal.

Definition 2.1 (Transportation Divergence between Mixtures) *Denote by $G = \sum_{i=1}^{I} w_i \{\theta_i\}$ and $\tilde{G} = \sum_{j=1}^{J} \tilde{w}_j \{\tilde{\theta}_j\}$ two discrete measures on space Θ. Let $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$ be their weight vectors, $f_i(x) = f(x|\theta_i)$, $\tilde{f}_j(x) = f(x|\tilde{\theta}_j)$, and $c(\cdot,\cdot): \mathcal{F} \times \mathcal{F} \to \mathbb{R}_+$ be a non-negative semi-continuous function. Let*

$$\Pi(\boldsymbol{w}, \tilde{\boldsymbol{w}}) = \{\boldsymbol{\pi} \in \mathbb{R}_+^{I \times J} : \boldsymbol{\pi} \mathbb{1}_J = \boldsymbol{w}, \boldsymbol{\pi}^\tau \mathbb{1}_I = \tilde{\boldsymbol{w}}\}.$$

Then the transportation divergence between two mixtures is given by

$$\mathcal{T}_c(f(\cdot\,|G), f(\cdot \mid \tilde{G})) = \inf \left\{ \sum_{i,j} \pi_{ij} c(f_i, \tilde{f}_j) \;\middle|\; \boldsymbol{\pi} \in \Pi(\mathbf{w}, \tilde{\mathbf{w}}) \right\}. \tag{2}$$

Unlike some divergence measures applied directly on two mixtures, the transportation divergence is computed through divergences between subpopulations. This choice greatly reduces the computational complexity and enhances the performance of the aggregated estimator. In particular, [1] develops an effective MM algorithm and recommends the KL divergence for the cost function $c(\cdot,\cdot)$. They further show that the resulting aggregated estimator is asymptotically consistent with a convergence rate $\mathrm{N}^{-1/2}$ when K is correctly specified. They also show that their algorithm for computing (2) always converges under mild conditions on the cost function. The implementation of this line of split-and-conquer approach does not rely on correctly specifying the order K. It is clear that the statistical performance of the aggregated estimator heavily depends on the knowledge of K. If K is under-specified, the estimators will be inconsistent. If the order is over-specified, it is interesting to know whether we can still obtain an aggregated estimator with adequate statistical performance. We aim to shed some light on this aspect of the approach through simulation study. We leave the potential remedies and other statistical issues as future work.

3. Simulations

We simulate data from a mixture of order K and have the simulated data partitioned into M subsets. On each local machine, we learn the mixture model based on the allocated subset under four scenarios. The first three scenarios are **i)** the order of the mixture is correctly specified; **ii)** the order of the mixture on the local machine is specified to be $K+1$; and **iii)** the order of the mixture on the local machine is specified to be $K+2$. In the fourth scenario, we learn a mixture of order $K+m$ on the mth local machine. In other words, the order of the mixture model varies with the local machines, and we refer to this scenario as mixed order hereafter. In the aggregation step, we combine the local estimates to form $\phi(\boldsymbol{x}|\bar{G})$ and reduce its order to the truth.

We study the performance of the GMR estimators with various choices of the divergence $D(\cdot\,|\,\cdot)$ in (1). The divergences considered in the experiment are:

1. **ISE**. The squared L_2 distance between two Gaussians, that is
$$\mathrm{D_{ISE}}\left(\phi(\cdot\,|\mathrm{G})|\phi\left(\cdot\,\middle|\widetilde{\mathrm{G}}\right)\right) = \left\|\phi(\cdot\,|\mathrm{G}) - \phi\left(\cdot\,\middle|\widetilde{\mathrm{G}}\right)\right\|_2^2 = \int\{\phi(x|\mathrm{G}) - \phi\left(x\middle|\widetilde{\mathrm{G}}\right)\}^2 dx.$$
2. **TD-KL**. The transportation divergence with the cost function being the KL divergence between two Gaussians in (2). This is the estimator considered in [1].
3. **TD-ISE**. The transportation divergence with the cost function being the squared L_2 distance between two Gaussians, which is $\mathrm{c}(\mathrm{f_i}, \tilde{f}_j) = \left\|\mathrm{f_i} - \tilde{f}_j\right\|_2^2$ in (2).

We also include a KL-averaging (KLA) approach of [10] in the simulation study. Unlike the above methods, KLA generates a sample of size 1000 from each locally learned $f\left(x\middle|\hat{G}_m\right)$ on a central machine. A maximum likelihood estimator (MLE) based on the pooled sample serves as the final aggregation result. Note that this approach does not require transferring the original data stored on local machines.

It can be shown that
$$\mathrm{D_{ISE}}\left(\phi(\cdot\,|\mathrm{G})|\phi\left(\cdot\,\middle|\widetilde{\mathrm{G}}\right)\right) = \boldsymbol{w}^{\tau} S_{GG}\boldsymbol{w} - 2\boldsymbol{w}^{\tau} S_{G\tilde{G}}\widetilde{\boldsymbol{w}} + \widetilde{\boldsymbol{w}^{\tau}} S_{\tilde{G}\tilde{G}}\widetilde{\boldsymbol{w}}$$
where $\mathrm{S_{GG}} = \{\phi\left(\boldsymbol{\mu}_{\mathrm{i}}\middle|\boldsymbol{\mu}_{\mathrm{j}}, \boldsymbol{\Sigma}_{\mathrm{i}} + \boldsymbol{\Sigma}_{\mathrm{j}}\right)\}$, $S_{G\tilde{G}} = \{\phi\left(\boldsymbol{\mu}_i\middle|\widetilde{\boldsymbol{\mu}}_j, \boldsymbol{\Sigma}_i + \widetilde{\boldsymbol{\Sigma}}_j\right)\}$, and $\mathrm{S_{\tilde{G}\tilde{G}}} = \{\phi\left(\widetilde{\boldsymbol{\mu}}_i\middle|\widetilde{\boldsymbol{\mu}}_j, \widetilde{\boldsymbol{\Sigma}}_i + \widetilde{\boldsymbol{\Sigma}}_j\right)\}$ are three matrices of shape $\mathrm{I}\times\mathrm{I}$, $I\times J$, and $J\times J$ respectively. With these expressions, computing $\mathrm{D_{ISE}}\left(\phi(\cdot\,|\mathrm{G})|\phi\left(\cdot\,\middle|\widetilde{\mathrm{G}}\right)\right)$ is an easy numerical task, but the minimization problem is harder to tackle. In our simulation, we use the BFGS algorithm [11] for optimization and the outcome may only be a local minimum.

At each local machine, we use the penalized MLE in [12] with the size of the penalty set to ${N_m}^{-1/2}$. When computing the KLA estimator on the central machine, we choose the penalty size by the same rule, which is $(1000M)^{-1/2}$. The EM-algorithm is used to compute pMLE and its convergence is claimed when the increment in the penalized log-likelihood function standardized by the total sample size is less than 10^{-6}. We use the *kmeans* algorithm to initialize the MM algorithm and we declare the convergence of the MM algorithm for the GMR estimator when the change in the objective function is less than 10^{-6}.

We generate data from the following four mixtures in our experiment. The first two are chosen as Gaussian mixtures of order $K=3$ and dimension $d=1$. Their density functions are given by

I. $\phi(x|\mathrm{G}) = 1/3\phi(x|{-3},1) + 1/3\phi(x|0,1) + 1/3\phi(x|3,1)$;

II. $\phi(x|\mathrm{G}) = 0.1\phi(x|{-2},1) + 0.3\phi(x|0,1) + 0.6\phi(x|3,1)$.

The next two are chosen as Gaussian mixtures of order $K=3$ and dimension $d=2$. To introduce the density function of these two mixtures, we first denote $\mu(\mathrm{r},\theta) = \mathrm{r}(\cos\theta, \sin\theta)^{\tau}$ and
$$\Sigma(\lambda_1,\lambda_2,\theta) = \begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}^{\tau}$$

III. $\phi(\boldsymbol{x}|\mathrm{G}) = 0.15\phi\left(\boldsymbol{x}\middle|\mu(2,3\pi/2),\Sigma(1{,}5{,}0)\right) + 0.35\phi\left(\boldsymbol{x}\middle|\mu(3,0),\Sigma(1{,}5,\pi/4)\right) + 0.5\phi\left(\boldsymbol{x}\middle|\mu(2,\pi/2),\Sigma(1{,}5,\pi)\right)$;

IV. $\phi(\boldsymbol{x}|\mathrm{G}) = 0.15\phi\left(\boldsymbol{x}\middle|\mu(2,3\pi/2),\Sigma(1{,}1{,}0)\right) + 0.35\phi\left(\boldsymbol{x}\middle|\mu(0,0),\Sigma(1{,}5,\pi/4)\right) + 0.5\phi\left(\boldsymbol{x}\middle|\mu(2,\pi/2),\Sigma(1{,}5,\pi)\right)$.

We generate samples of sizes $\mathrm{N}=2^{19}$ or $N=2^{21}$ respectively from each of the mixtures given above. Each sample is then split into $M=4$ or $M=8$ subsets completely at random and they are regarded as stored on $M=4$ or $M=8$ local machines. With these two choices of sample sizes and two choices of the number of local machines, we obtain 4

combinations. The four split-and-conquer methods discussed earlier are applied to obtain the aggregated estimates. We assess the performances of these 4 methods based on $R = 100$ repetitions with the following metrics.

a. **ISE**. We compute the ISE between the estimated mixture and the true mixture.
b. **Adjusted Rand Index (ARI)**. The finite mixture model is often used for model-based clustering. We may divide the observations into K clusters based on the true mixture or the learned mixture. An ARI value proposed by [13] measures the similarity of two clustering outcomes.
c. **Computational time**. One important consideration of a split-and-conquer approach is the amount of computation. A conceptual optimal approach is not useful if it cannot be solved within a reasonable amount of time.

3.1. Simulation Results under Distributions I and II

We summarize the results in Figure 1 when the split-and-conquer methods are applied to data generated from distributions I and II. Note the plots in the first and second columns are results under distributions I and II respectively. We then divide each plot into 4 panels labelled by $M = 4$ or $M = 8$ on the top, and $N = 2^{19}$ or $N = 2^{21}$ on the right margin with an obvious interpretation. Within each panel are boxplots of one of the performance measures for 4 methods. The lower ISE and higher ARI indicate better performance.

According to Figure 1, when the true order $K = 3$ is specified at local machines, ISE, TD-ISE, and TD-KL have similar and good performances. When the number of machines increases or the sample size increases, the boxplots get shorter, indicating lower variations. In comparison, the ISE of KLA approach is hundreds of times larger. It is therefore less efficient.

When the order on local machines is over-specified with $K = 4$, the ISE method is negatively but only mildly affected in terms of both ISE and ARI. The TD-ISE and TD-KL become much worse and less stable. The KLA remains non-competitive. When the order is over-specified at $K = 5$, the ISE remains well behaved. The computationally favoured TD-ISE and TD-KL become statistically ineffective. Under the case of mixed orders, the ISE is still well behaved and other methods remain non-competitive.

3.2. Simulation Results under Distributions III and IV

The simulation results under distributions III and IV are summarized in Figure 2. The plots in Figure 2 are arranged the same way as before. The plots in the first and second columns are results under distributions IIII and IV respectively.

Most statistical methods have deteriorated performance under multi-dimensional data. It turns out that the performance of the ISE approach remains reasonable in all cases under distribution III. When the order is slightly over-specified with $K = 4$, the performances of TD-ISE and TD-KL also remain reasonable. When $K = 5$, these two approaches become unstable, just like their performance under distributions I and II. Under the case of mixed orders, the ISE approach is still well behaved. The other methods remain non-competitive.

Unlike distributions I-III, the subpopulations in distribution IV are not well separated. We anticipate that all approaches do not perform too well. Indeed, ISE, TD-ISE, and TD-KL are all unstable even when the order is correctly specified with $K = 3$. We are surprised, however, that they recover from this failure when the order is over-specified at $K = 4$. The ISE approach has a comparable low ISE value in this case to the ISE value under distribution III, where the subpopulations are well separated.

3.3. Summary

Our simulation experiment only covers a small range of scenarios. It is dangerous to generalize what we have observed. It might be safe to say, the reasonable performance of the ISE approach in all situations included indicates the local estimates effectively summarize the information contained in the data. The over-specification of the order may not always be devastating. Based on these results, one may decide to always use the ISE approach as it is least affected by over-specification. However, the drawback of this approach is its computational complexity. It becomes infeasible when either M or the dimension d becomes larger.

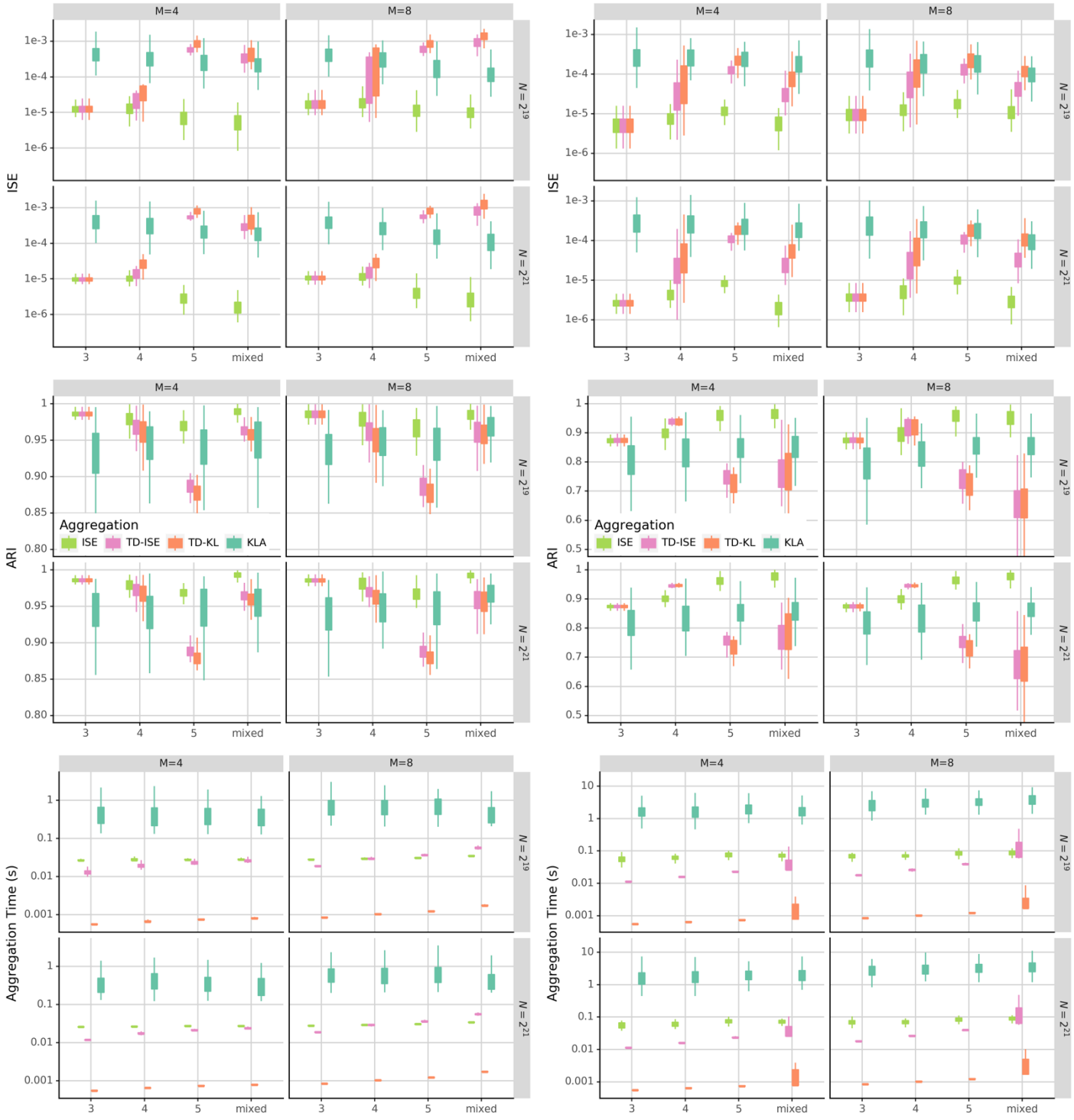

Fig. 1: Performances of four split-and-conquer approaches for learning 1-dimensional 3-component mixtures.

4. Conclusion

In this paper, we investigate the potential adverse effect of over-specifying the order of the finite mixture model at local machines in split-and-conquer approaches. The scale of our experiment is limited and does not cover sufficient grounds. We only simulate data from mixtures in low dimensional space and with relative low order. Surprisingly, the simple ISE approach has very good statistical performance and it is robust to order over-specification in general. The straightforward

implementations of TD-ISE and TD-KL do not perform as well under order misspecification. One is reminded that the motivation behind the TD-ISE and TD-KL is their computational efficacy, which is not shared by ISE. The superior performance of ISE in terms of the robustness against over-specification indicates the statistical efficacy is possible. With some effort, we believe robust as well as computationally efficient split-and-conquer approaches can be found. We aim to pursue this topic in the future.

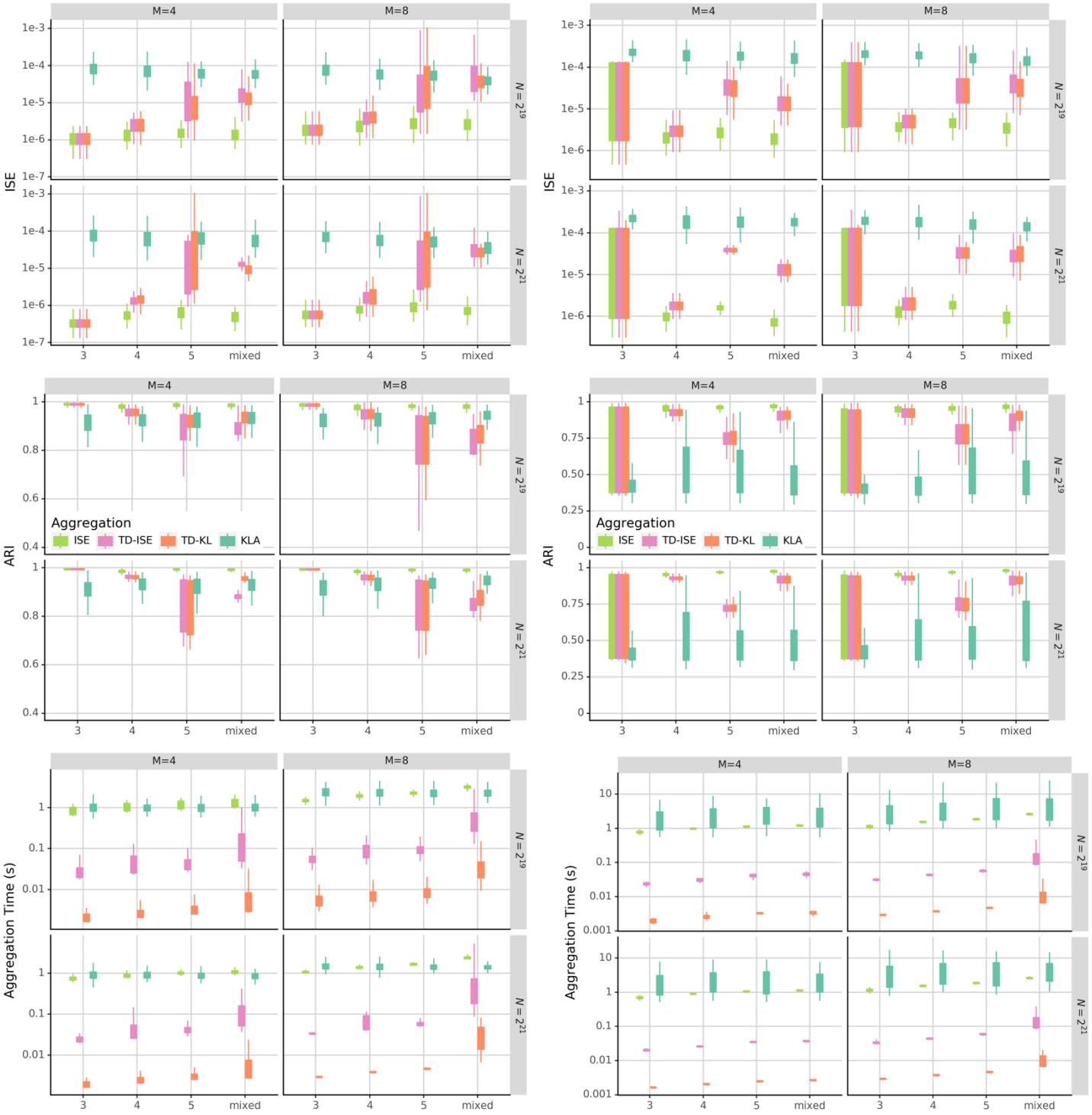

Fig. 2: Performances of four split-and-conquer approaches for learning 2-dimensional 3-component mixtures.

References

[1] Q. Zhang and J. Chen, "Distributed Learning of Finite Gaussian Mixtures," *ArXiv201010412 Stat*, Feb. 2021, Accessed: Jul. 14, 2021. [Online]. Available: http://arxiv.org/abs/2010.10412

[2] X. Chen and M. Xie, "A split-and-conquer approach for analysis of extradinarily large data," *Stat. Sin.*, vol. 24, no. 4, pp. 1655–1684, 2014.

[3] Y. Zhang, J. Duchi, and M. Wainwright, "Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm with Minimax Optimal Rates," *J. Mach. Learn. Res.*, vol. 16, no. 102, pp. 3299–3340, 2015.

[4] X. Chang, S.-B. Lin, and Y. Wang, "Divide and conquer local average regression," *Eletronic J. Stat.*, vol. 11, no. 1, pp. 1326–1350, 2017.

[5] D. F. Crouse, P. Willett, K. Pattipati, and L. Svensson, "A look at Gaussian mixture reduction algorithms," Chicago, IL, USA, 2011, pp. 1–8. Accessed: Jul. 14, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/5977695

[6] D. Schieferdecker and M. F. Huber, "Gaussian mixture reduction via clustering," in *2009 12th International Conference on Information Fusion*, Jul. 2009, pp. 1536–1543.

[7] J. L. Williams and P. S. Maybeck, "Cost-function-based hypothesis control techniques for multiple hypothesis tracking," *Math. Comput. Model.*, vol. 43, no. 9, pp. 976–989, May 2006, doi: 10.1016/j.mcm.2005.05.022.

[8] L. Yu, T. Yang, and A. B. Chan, "Density-Preserving Hierarchical EM Algorithm: Simplifying Gaussian Mixture Models for Approximate Inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1323–1337, Jun. 2019, doi: 10.1109/TPAMI.2018.2845371.

[9] Q. Zhang and J. Chen, "A Unified Framework for Gaussian Mixture Reduction with Composite Transportation Distance," *ArXiv200208410 Cs Stat*, Feb. 2020, Accessed: Jul. 14, 2021. [Online]. Available: http://arxiv.org/abs/2002.08410

[10] Q. Liu and A. Ihler, "Distributed Estimation, Information Loss and Exponential Families," 2014, pp. 1098–1106.

[11] J. Nocedal and S. J. Wright, Eds., "Quasi-Newton Methods," in *Numerical Optimization*, New York, NY: Springer, 2006, pp. 135–163. doi: 10.1007/978-0-387-40065-5_6.

[12] J. Chen and X. Tan, "Inference for multivariate normal mixtures," *J. Multivar. Anal.*, vol. 100, no. 7, pp. 1367–1383, Aug. 2009, doi: 10.1016/j.jmva.2008.12.005.

[13] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *J. Am. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971, doi: 10.2307/2284239.