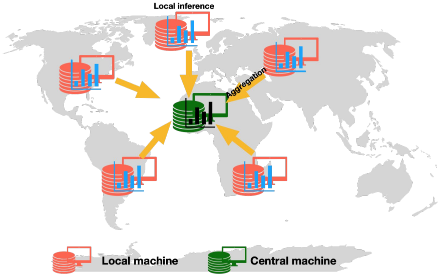
Department of Statistics, University of British Columbia

Distributed Learning of Finite Gaussian Mixtures

Qiong Zhang* and Jiahua Chen *presenter

Background

- Distributed learning: datasets are too large to be stored on a single facility or collected by different agencies and cannot be shared due to privacy.
- Split-and-conquer (SC) approach:
  - ○ Local inference: standard inference on local machine;
  - ○ Aggregation: combine local results to a central machine.



- Advantages of split-and-conquer:
  - ○ Only share summary Statistics to address privacy concern
  - ○ Only needs one round of communication
- Finite Gaussian mixture models (GMM)
  - ○ Density of GMM: $\phi_G(x) := \int \phi(x;\theta)dG(\theta) = \sum_{k=1}^{K} w_k\phi(x;\theta_k)$
  - ○ Mixing distribution: $G = \sum_{k=1}^{K} w_k\delta_{\theta_k}$
  - ○ Order of a GMM: K (assumed to be known).
  - ○ Parameter space of order K mixture: mixing distribution with up to K support points.

Challenge

- Aggregation when model parameter space is Euclidean: simple average.
- Aggregation under GMM:
  - ○ Simple average: $\bar{G} = \sum_{m=1}^{M} \lambda_m \hat{G}_m$
  - ○ The simple average usually has MK components, and is **NOT** in the parameter space
  - ○ Average mixture: $\phi_{\bar{G}}(x) = \sum_{m=1}^{M} \lambda_m \phi_{\hat{G}_m}(x)$
  - ○ The average mixture is the good estimate for the true mixture.

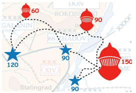
Proposed Method

- Big picture: approximate the average mixture with a mixture with order K in a ***statistically*** and ***computationally*** efficient way.
- Reduction estimator

$$\bar{G}^R = \mathrm{argmin}_{G\in\mathbb{G}_K}\rho(\Phi_{\bar{G}}, \Phi_G)$$

where the divergence is chosen to be the composite transportation divergence (CTD).

- Composite transportation divergence
  - ○ Byproduct of optimal transportation.
  - ○ View GMMs as discrete distribution on the space of Gaussian distributions.
  - ○ Let $\Phi_G(x) = \sum_{n=1}^{N} w_n\Phi(x;\theta_n) := \sum_{n=1}^{N} w_n\Phi_n$ and $\Phi_{G'}(x) := \sum_{m=1}^{M} w'_m\Phi'_m$, let $c(\cdot,\cdot)$ be a non-negative function, the CTD between two mixtures is defined as



$$\mathcal{T}_c(\Phi_G, \Phi_{G'}) = \min\left\{\sum_{n,m}\pi_{nm}c(\Phi_n, \Phi'_m) : \sum_{m}\pi_{nm} = w_n, \sum_{n}\pi_{nm} = w'_m\right\}.$$

Numerical Computation

- Equivalent simplified optimization.
  - ○ The simplified objective function has a closed-form.
  - ○ The optimization over mixing weights and subpopulations is separated.

$$\mathcal{J}_c(\Phi_{\bar{G}}, \Phi_{G'}) = \min\left\{\sum_{n,m}\pi_{nm}c(\Phi_n, \Phi'_m) : \sum_{m}\pi_{nm} = w_n\right\}$$

$$\bar{G}^R = \mathrm{arginf}_{G\in\mathbb{G}_K}\mathcal{J}_c(\Phi_{\bar{G}}, \Phi_G) \quad \bar{w}_j^R = \sum_i \pi_{ij}^*$$

- Efficient MM algorithm.
- Can be viewed as a k-means algorithm in the space of Gaussian distributions.

Statistical Property

Theorem

C1 The data are IID observations from $\Phi_{G^*}(x)$ with order K.

C5 Local triangular inequality
$A^{-1}\|\Phi_1 - \Phi_2\|^2 \le c(\Phi_1, \Phi_2) \le A\|\Phi_1 - \Phi_2\|^2.$

Under conditions C1-C5, up to permutations, we have

$$\bar{\Phi}^R - \Phi_k^* = O_p(N^{-1/2}), \quad \bar{w}^R - w_k^* = O_p(N^{-1/2}).$$

Simulation Study

The setting of the simulation study is described as follows.

Setting

- Generate 100 random GMM in 50-dimension ￼with K=5￼.
- MaxOmega: degree of overlap between subpopulation is set to be 1%, 5%, 10%.
- Total sample size ￼N=2^{21}.
- The number of local machines are set to M=4, 16, 64￼.

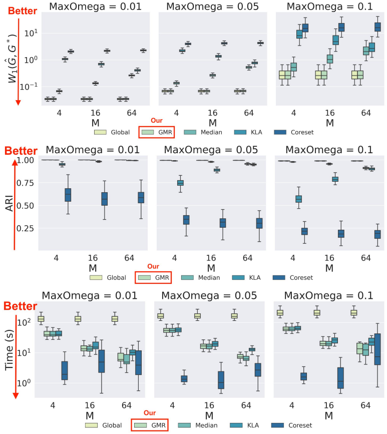
Estimators for Comparison

- **Global**: the estimator based on the full dataset.
- **Median**: the “best” local estimator.
- **GMR**: our method with KL divergence as cost function.
- **KLA** in [1].
- **Coreset** in [2].

Results

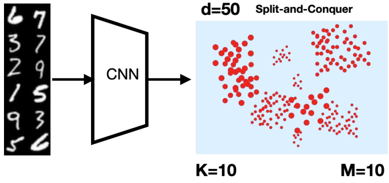
Performance of Estimators in Simulation

Comparison of Global, Median, GMR, KLA, and Coreset estimators for learning 50-dimensional 5-component mixture when sample size is 2^{21} in terms of the statistical efficiency and computational efficiency.



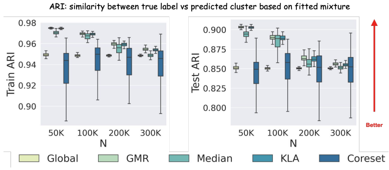
NIST Clustering

- Apply the proposed method on NIST dataset [3] for the handwritten character recognition.
- Use a pre-trained CNN to extract image feature to d=50.
- Fit a 10-component GMM: the feature for each digit forms its own subpopulation.



Performance of Estimators on Real Dataset

Training and test ARI of Global, Median, GMR, KLA, and Coreset estimators for learning of 10-component GMM on 50-dimensional space for NIST digit classification on 10 machines.



Reference

1. Liu, Q. and Ihler, A., 2014. Distributed estimation, information loss and exponential families. In 2014 Advances in Neural Information Processing Systems 27, pages 1098–1106.
2. Lucic, M., Faulkner, M., Krause, A. and Feldman, D., 2017. Training Gaussian mixture models at scale via coresets. The Journal of Machine Learning Research, 18(1), pp.5885-5909.
3. P. Grother and K. Hanaoka. NIST special database 19 handprinted forms and characters 2nd edition. Technical report, National Institute of Standards and Technology, 2016.

Acknowledgement

This research was enabled in part by support provided by WestGrid (www.westgrid.ca) and Compute Canada (www.computecanada.ca).

THE UNIVERSITY OF BRITISH COLUMBIA

Nov 20, 2021 @ Gather.town

CANSSI INCASS

Canadian Statistical Sciences Institute
Institut canadien des sciences statistiques